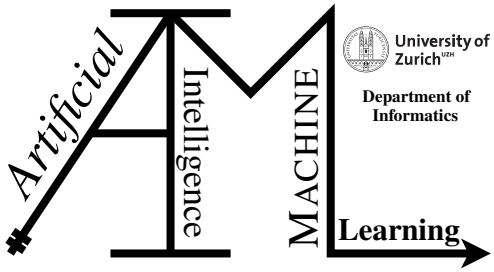
MR-based synthetic CT generation for MR-guided radiotherapy

Master Thesis

Mariia Lapaeva
18-747-915

Submitted on
April 01 2022

Thesis Supervisors
Prof. Dr. Manuel Günther, Agustina La Greca Saint-Esteven, Dr. Riccardo Dal Bello, Dr. Stephanie Tanadini-Lang, Prof. Dr. Ender Konukoglu



Master Thesis

Author: Mariia Lapaeva, mariyalapaevan@gmail.com

Project period: 01.10.2021 - 01.04.2022

Artificial Intelligence and Machine Learning Group
Department of Informatics, University of Zurich

Acknowledgements

First of all, I would like to thank a lot my supervisors, who have supported me from the beginning and provided me with the opportunity to work on the interdisciplinary project I have been dreaming of.

I am very grateful to Agustina La Greca for her imaginative ideas that enabled this project to come to life, her thorough guidance, eternal inspiration and tireless faith. I would like to thank Riccardo Dal Bello for his in-depth expertise, always positive attitude and, especially, for making my immersion in the field of radiology very enjoyable. I appreciate all the discussions we had together in our small group, which showed me a direction for further advancement.

With Prof. Manuel Günther's supervision, I always felt supported and empowered, which enabled me to focus better on the subject. I am grateful to Prof. Günther for his very helpful guidance on general topics and finer details, and for his fresh perspective on medical imaging, which allowed me to move forward with the project.

I would like to thank Dr. Stephanie Tanadini-Lang for her kindness, careful guidance, the detailed feedback that she gave me during the work on the thesis and the scientific research inspiration she bestowed upon me.

I am very grateful to Prof. Ender Konukoglu for the inspiring lectures that have reinforced my interest in medical image analysis, his valuable support and the computational resources made available to me.

I very much appreciate an extraordinarily encouraging atmosphere at UZH, USZ and ETHZ and I am very grateful to all the lecturers and students with whom I worked during my studies. Especially, I would like to thank Philipp Wallimann for all of the discussion we had about intensity normalisation and machine learning, and for his very kind and cooperative support.

Finally, I would like to thank my family and friends, without whom this would not be possible, who are always there, despite the time of day, distance and circumstances, for their faith, support, trust, humour and for being my source of inspiration.

Abstract

The aim of this study is to devise deep learning (DL) approaches trained on paired and unpaired data that are able to generate realistic synthetic computed tomography (CT) images for magnetic resonance-guided radiotherapy (MRgRT) in the area of the abdomen and to assess its clinical applicability. The imaging data of 76 patients with a tumour in the abdomen who were treated with MRgRT at USZ was collected retrospectively and divided in training (59) and test sets (17). To improve the current state-of-the-art of DL technologies by studying different architectures and ensembles of configurations, the following experiments were conducted: (a) evaluating the influence of the different GAN architectures trained on paired (Pix2pix) and unpaired data (CycleGAN and CUT, which firstly applied for the purpose of sCT generation); (b) investigating the footprint of different preprocessing methods (Nyul, novel N-peaks); (c) improving spatial consistency of results by adjusting the network configuration (2D, pseudo3D); (d) testing different GAN training objectives (LSGAN, WGAN-GP); (e) estimating the influence of the loss function on the generated results (per-pixel L1 loss, VGG19 perceptual loss). The quality of sCT generation was assessed using both, image similarity and dosimetric accuracy metrics. The dosimetric accuracy of the best performing models was estimated by comparing the dose distribution of MRgRT treatment plans calculated from synthetic CT and original CT images using dose-volume histogram (DVH) parameters to allow assessment of the clinical applicability of the DL methods. Our results suggest that DL models trained with unpaired data achieve similar performance as models requiring perfectly aligned image pairs, and even perform better in the bone and air pocket areas. The mean absolute errors (mean ± SD) calculated within the body contour are 71.0±20, 73.4±21 and 84.5 ±19 HU when using the best performing configuration of pix2pix, CycleGAN and CUT, respectively. The proposed DL-based synthetic CT generation methods may be considered clinically applicable for treatment planning in the abdominal area with the mean DVH indicator discrepancies with the original plan of less than 1% for all models, and less than 0.5% for all tumour DVH indicators for the best performing model, CycleGAN. The study confirmed that generation of synthetic CT using a DL approach from low field magnetic resonance images in the abdomen is feasible and allows a reliable calculation of irradiation plans in MRgRT.

Zusammenfassung

Das Hauptziel dieser Studie ist die Entwicklung von Deep Learning (DL) Ansätzen, die auf gepaarten und ungepaarten Daten trainiert werden und in der Lage sind, realistische synthetische Computertomographie (CT)-Bilder für die magnetresonanzgeführte Strahlentherapie (MRgRT) im Abdomen zu generieren und ihre klinische Anwendbarkeit zu evaluieren. Die Bildgebungsdaten von 76 Patienten mit einem Tumor im Abdomen, die am USZ mit MRgRT behandelt wurden, wurden retrospektiv gesammelt und in eine Trainings- (59) und Testgruppe (17) unterteilt. Um den aktuellen Stand der DL-Techniken durch die Untersuchung verschiedener Architekturen und Ensembles von Konfigurationen zu verbessern, wurden die folgenden Experimente durchgeführt: (a) Bewertung des Einflusses der verschiedenen GAN-Architekturen, die auf gepaarten (Pix2pix) und ungepaarten Daten trainiert wurden (CycleGAN und CUT, die erstmals für diesen Zweck der sCT-Generierung eingesetzt wurden); (b) Untersuchung des Einflusses verschiedener Vorverarbeitungsmethoden (Nyul, neue Methode N-Peaks); (c) Verbesserung der räumlichen Konsistenz der Ergebnisse durch Anpassung der Netzwerkkonfiguration (2D, pseudo3D); (d) Test der verschiedenen GAN-Trainingsziele (LSGAN, WGAN-GP); (e) Abschätzung des Einflusses der Verlustfunktion auf die generierten Ergebnisse (L1-Fehler pro Pixel, VGG19-Wahrnehmungsfehler). Die Qualität der sCT-Generierung wurde sowohl anhand der Bildähnlichkeit als auch der dosimetrischen Genauigkeit bewertet. Die dosimetrische Genauigkeit der leistungsstärksten Modelle wurde durch den Vergleich der Dosisverteilung von MRgRT -Behandlungsplänen geschätzt, die anhand von synthetischen CT- und Original-CT-Bildern unter Verwendung von Dosis-Volumen-Histogramm-Parametern (DVH) berechnet wurden, um die klinische Anwendbarkeit der DL-Methoden zu bewerten. Unsere Ergebnisse deuten darauf hin, dass DL-Modelle, die auf ungepaarte Weise trainiert wurden, eine ähnliche Leistung erzielen wie Modelle, die perfekt übereinstimmende Bildpaare erfordern, und in den Knochen- und Luftblasenregionen sogar besser abschneiden. Die Mittleren Absoluten Fehler (Mittelwert ± Standardabweichung), die innerhalb der Körperkontur mit der leistungsstärksten Konfiguration von pix2pix, CycleGAN und CUT berechnet wurde, betrug jeweils 71.0±20, 73.4±20 und 84.5 ±19 HU. Die vorgeschlagenen DL-basierten Generierungsmethoden synthetischer CT können als klinisch ausreichend für die Behandlungsplanung bei Tumoren im Abdomen angesehen werden, da die mittleren Abweichungen der DVH-Indikatoren vom ursprünglichen Plan für alle Modelle weniger als 1% betragen. Für das leistungsstärkste Modell, CycleGAN, betragen alle Tumor-DVH-Indikatoren sogar weniger als 0.5%. Diese Studie hat bestätigt, dass die Generierung synthetischer CT mit Hilfe eines DL-Ansatzes für Magnetresonanzbilder (MR) des Abdomens mit schwachem Magnetfeld machbar ist und eine zuverlässige Berechnung von Bestrahlungsplänen in der MRgRT ermöglicht.

Contents

Chapter 1

Introduction

1.1 Motivation

A leading cause of death worldwide is cancer, accounting for about 10 million deaths in 2020, or nearly one in six deaths, according to the World Health Organisation, WHO (2022). Cancer treatment modalities is a combination of radiotherapy, surgery, chemotherapy, immunotherapy and hormone therapy. Radiotherapy remains an integral part of cancer treatment: about 50% of all cancer patients receive radiotherapy during the course of their disease; it contributes to 40% of curative treatment of cancer (Baskar et al., 2012). Radiation therapy aims to deliver a dose of radiation that destroys the target tumour cells and spares healthy tissue from radiation. To achieve this, magnetic resonance-guided radiation therapy (MRgRT) was intriduced, which is one of the latest technical achievements in the field of radiation oncology. In MRgRT, treatment and imaging are performed simultaneously with a hybrid device composed of a linear accelerator (LINAC) coupled to a magnetic resonance scanner. Compared to standard methods, in MRgRT exposure to radiation only occurs when the detected tumour is in the correct position in the cine magnetic resonance (MR) image, and the treatment beam switches off automatically when the tumour moves outside a certain boundary, as shown in Figure 1.1. This keeps radiation to unwanted areas to a minimum and allows the therapy to be applied.

In practice, MR images on the MR Linac are used to set up the patient and adapt the treatment to the daily anatomy without ionising radiation. MR images are characterised by absense of ionizing radiation and a high soft tissue contrast, which motivates their clinical application in delineating tumours and their target volumes for irradiation, as well as organs at risk (OAR) whose irradiation could cause damage. In principle, MR signal can be interpreted as a map of density of water protons in the tissues of the body. However, the intensity on MR images is a relative measurement, which could vary due to different imaging settings, devices and even the presence of implants. This does not provide a direct estimate of tissue density, more in particular electron density (ED), which is required to calculate the radiation dose.

This is why computed tomography (CT) plays a fundamental role in radiotherapy. The CT scan is used to directly estimate the electron density map by precise calibration of Hounsfield units (HU) and consequently applied for calculating the dose distribution. The drawbacks of CT include radiation exposure to the patient, limited soft tissue contrast, additional burden on patients and cost for the healthcare system. Therefore, in today's medical practice in MRgRT, CT scans are only taken during the planning phase. During treatment sessions, the benefits of complementary modalities are utilised: current standard practice includes the generation of synthetic CT (sCT) at treatment visits for dose estimation, based on the deformable registration of the original CT with the MR of the day. This may however introduce additional alignment uncertainties, for example due to changes in anatomy over time, leading to errors of up to 3 mm and an increase

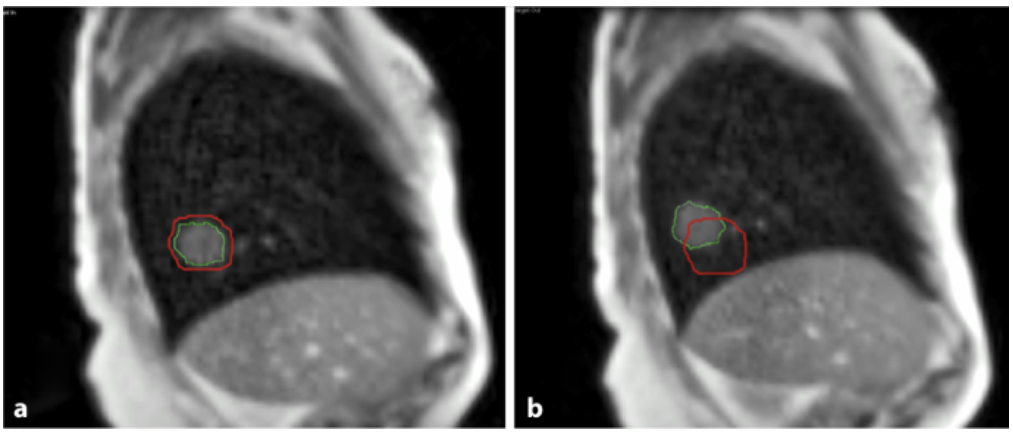


Figure 1.1: Automated beam gating based on cine-magnetic resonance imaging in a sagittal plane through the tumor. Cine MR images provide detailed information about both the anatomy and the dynamic movement of the airways. Beam is on while the target cancer volume (green) is within bounds (red; a) and automatically turned off when more then predefined fraction of the target is outside the defined boundary (b). Source: Spindeldreier et al. (2021)

in the target volume. This would, in turn, affect the accuracy of the overall treatment and would mean damage to healthy tissue (Spadea et al., 2021; Ulin et al., 2010; Nyholm et al., 2009).

To overcome these limitations, several approaches have been proposed. Among these are the bulk density and multi-atlas approaches. The bulk density approach, which employs a rigid image registration technique combined with a manual outer body correction scheme, has provided fairly efficient and accurate results. However, it requires manual electron density assignment to delineated tissues on MR images. Apart from that, it requires costly expert labour. That is why multi-atlas approaches, which are based on the deformable registration of image sets (atlases) previously delineated by trained experts, have been introduced recently. The aim of the atlas-driven approach is to encode the relationship between the segmentation labels and the image intensities observed in the atlases in order to assign segmentation labels to the voxels of an unlabelled image (Pham et al., 2000). To obtain the sCT of a treatment visit, the spatial correspondence between a patient's unlabelled MR volume and an MR atlas is established, and the labels are further propagated. The calculated deformation map is then applied to the CT scans of the atlas. Due to the high computational complexity of deformable registrations, the multi-atlas approach is time-consuming (Iglesias and Sabuncu, 2015; Demol et al., 2016; Andres et al., 2020).

Time and accuracy are key factors in cancer treatment that require further exploration and development of the currently proposed methods, as adaptive MRgRT involves the patient waiting in the treatment position until the plan adaptation is complete. One of the most promising methods is a deep learning (DL) based generation of sCTs. Neural networks (NN) have demonstrated their potential in various image-to-image translation tasks, including the generation of sCT from MR images. Research in the last three years showed that Generative adversarial networks (GANs), which employ more than one loss in training, could outperform traditional convolutional neural networks (CNN) architectures, which compute a single loss function between input and output images for optimisation. GANs have become the state-of-art technique for the sCT generation task (Spadea et al., 2021). At present, the most commonly used architectures are Image-to-Image Conditional Generative Adversarial Network (Pix2Pix), introduced by Isola et al. (2017), and Cycle-consistent Generative Adversarial Network (CycleGAN), introduced by Zhu et al. (2017). While the Pix2Pix model requires co-registered image pairs, CycleGAN only

requires images from each modality rather than nearly perfectly aligned image pairs of the same patient. Due to the scarcity of data, the importance of architectures that do not require the alignment of image pairs is increasing dramatically. One of the latest models developed by Park et al. (2020) is Contrastive Learning for Unpaired Image-to-Image Translation (CUT). The model uses a patch-based approach instead of manipulating the content of an entire image. Corresponding patches in the input and output image should have high mutual information. The proposed framework enables one-sided translation without image alignment. It was successfully applied to the different image-to-image translation tasks, including style transfer, object transfiguration, season transfer and photo enhancement (Han et al., 2021). To the best awareness, CUT has not yet been used for the task of synthetic CT generation in the abdomen region, which motivates the current work.

Most DL approaches for sCT generation are based on paired data, i.e. intrinsically registered CT and MR scans acquired on the same day and at the same patient position. These conditions are scarce in the currently available public datasets of medical images as well as in our partner institution, the University Hospital Zurich (USZ). Fewer studies are available for the abdominal field of view (FOV) because the study population is more difficult due to organ movement, the presence of air pockets and co-registration issues.

1.2 Goals

The aim of this project is to devise deep learning approaches trained on paired and unpaired data that are able to generate realistic sCT images from MR images. The proposed method intends to improve current state-of-the-art DL techniques by studying different architectures and ensembles thereof, as well as different combinations of preprocessing methods and losses best suited for the proposed task. To assess the clinical applicability of the method, the accuracy of sCT-based dose calculation will be evaluated.

The objectives of this project are formulated in the following research questions:

$\boldsymbol{RQ1}$: Could NN architectures, trained in unpaired fashion, achieve similar performance as architectures, requiring perfectly aligned image pairs in the abdominal area?

$\boldsymbol{RQ2}$: Could biologically motivated normalisation methods improve the performance of NNs for sCT generation by focusing on specific tissue intensity correction?

$\boldsymbol{RQ3}$: Could a NN trained with the help of three adjacent 2D slices avoid 3D discontinuities in the area of the abdomen, which is heavily affected by respiratory and peristaltic changes?

$\boldsymbol{RQ4}$: Could different GAN objectives by improving the optimisation process result in a better quality of generated sCTs?

$\boldsymbol{RQ5}$: Would using a perceptual loss function in generator instead of a per-pixel loss function help to overcome the known problems in abdomen sCT generation: fuzzy organ boundaries and bone formation errors?

1.3 Structure

Chapter 2 provides the clinical background for MR-guided radiotherapy. Chapter 3 goes into deeper detail of the current Deep Learning based techniques for image-to-image translation tasks, including the sCT generation. In Chapter 4, we present the different experiments which aim to improve the current state-of-the-art DL approaches for sCT generation. This includes the image acquisition and preprocessing steps, the set of applied NN architectures, their modifications, and various loss functions. Chapter 5 presents the achieved results, which were assessed by means of both image similarity and task-specific (dosimetry accuracy) metrics. Chapter 6 provides the

discussion on the experiment results and threats to validity. Chapter 7 is a general conclusion to the work.

Chapter 2

Background: Medical Imaging In Radiotherapy

2.1 Magnetic Resonance Imaging

MR imaging is a medical imaging technique that uses a magnetic field and radio frequency (RF) signals to produce images of anatomical structures. The only substance in tissue with a sufficient concentration of magnetic nuclei to produce good images is hydrogen in water molecules. The nucleus of a hydrogen atom consists of a single proton. Therefore, the MR image is an image of water protons. When tissue containing hydrogen (small magnetic nuclei), i.e. protons, is placed in a strong magnetic field, some protons orient themselves in the same direction as the magnetic field. MR imaging requires a magnetic field that is both strong and uniform. The field strength of the magnet is measured in teslas (T). The standard field is 1.5T, which can go up to 3T for the brain imaging, positively affecting the quality of alignment. This alignment creates magnetisation in the tissue, which in a following relation process generates the RF signal detectable by the MR scanner's receiver coils. These signals are used to create MR images of the body. A brief overview of this imaging process is demonstrated in Figure 2.1. If a tissue does not have a sufficient concentration of hydrogen-containing molecules, it is not visible in the MR image. On MR images, soft tissues with a higher hydrogen concentration, such as fat, muscles or tumours, are clearly discernible, while bone structures are barely identifiable.

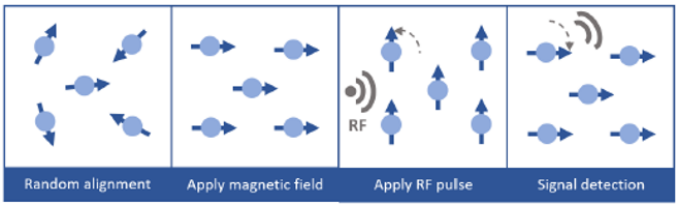


Figure 2.1: Simplistic overview of how protons are used for measuring a MRI signal. When an RF current is pulsed through the patient, the protons are stimulated, and spin out of equilibrium, straining against the pull of the magnetic field. When the radiofrequency field is then turned off, the MRI sensors are able to detect the energy released as the protons realign with the magnetic field. The time it takes for the protons to realign with the magnetic field, as well as the amount of energy released, changes depending on the environment and the chemical nature of the molecules. Physicians are able to tell the difference between various types of tissues based on these magnetic properties. Sources: Rosbergen (2021); NIBIB (2022)

MR imaging produces images that are different from images produced by other imaging modalities. An advantage of MRI is the ability to selectively image a variety of tissue by selecting the appropriate RF sequence. If a certain pathological condition is not visible in one sequence, it is possible to detect it with a different sequence. During the imaging procedure, an image recording of the patient's body is first divided into slices, and then slices are divided into a matrix of voxels. Each voxel is an independent RF signal source. The voxel size can be adjusted, in turn determining the detail of the image and also affecting the image noise. The five most important image quality characteristics - contrast sensitivity, detail, noise, artefacts and spatial resolution - can be largely controlled by the settings of the various protocol factors, as described by Sprawls (2000). MR imaging is done without ionising radiation, so patients are not exposed to the harmful effects of ionising radiation.

2.2 Computed Tomography Imaging

Computed tomography is a common imaging technique as it overcomes most of the limitations of traditional radiography. In particular, CT allows the differentiation of tissues with very narrow attenuation coefficients and the visualisation of 3D volumes with a high resolution.

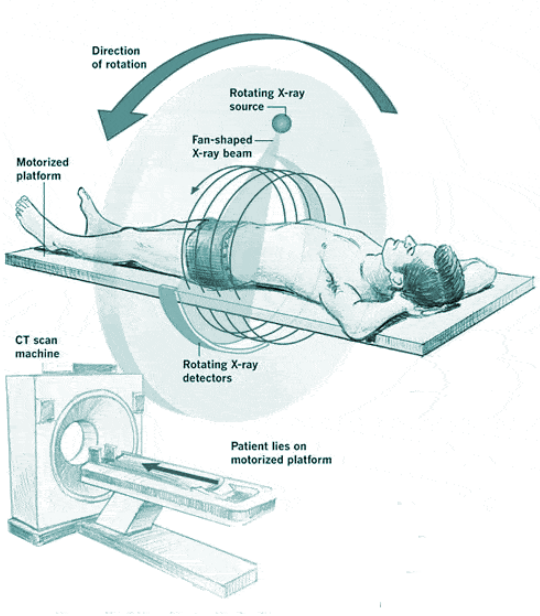


Figure 2.2: Simplistic overview of CT imaging process. Source: Ahad (2015)

The principle of CT imaging process is as follows: multiple projections are taken with an X-ray tube that can produce an X-ray beam and a group of detectors that rotate around the patient, the arrangement and number of which can vary according to the generation of the device (see Figure 2.2). Those projections are further reconstructed into a 3D image. On these images the density of different biological tissues is calculated using a linear attenuation coefficient μ expressed in Hounsfield units (HU):

$$\mu(HU) = 1000\frac{\mu - \mu(H_2O)}{\mu(H_2O)} \tag{2.1}$$

where μ represents the linear attenuation coefficient of the hit tissue, while $\mu(H_2O)$ represents the linear attenuation coefficient of the water. The principle for calculating the attenuation coefficients is based on Lambert-Beer's law, which relates the change in the number of photons after hitting a material to a certain linear attenuation coefficient:

$$N = N_0 e^{-\mu x} \tag{2.2}$$

where N_0 represents the number of photons striking the material, while N represents the attenuation after they have passed through it. The attenuation coefficient is indicated by μ, and x represents the space covered in the material.

On CT images, the Hounsfield scale is shown in greyscale. Denser tissue with greater X-ray absorption has positive values and appears bright; less dense tissue with lower X-ray absorption has negative values and appears dark (Chappell, 2019; Broder and Preston, 2011). An example of the HU scale of a head CT image is shown in Figure 2.3.

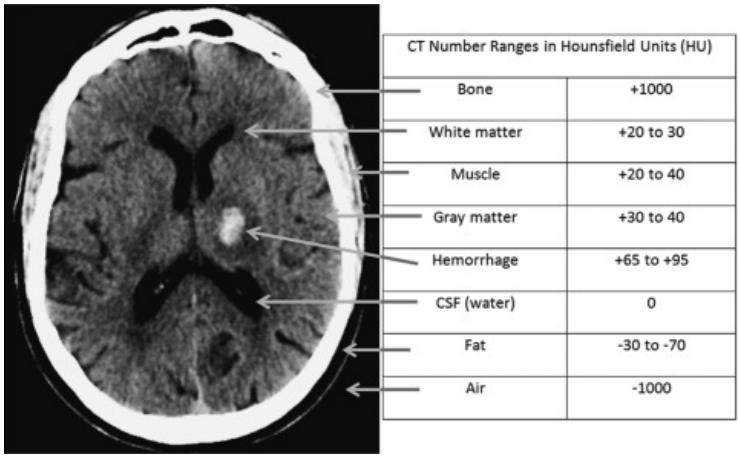


Figure 2.3: Approximate HU values for tissues commonly found on head CT images. Source: Kamalian et al. (2016)

On the whole, CT images are characterised by very accurate spatial information, information on tissue density required for dosimetry calculations, and distinguishable bones that enable the detection of bone tumour. The main disadvantages are the rather suboptimal imaging of soft tissues and the ionising radiation of healthy tissues.

2.3 Magnetic Resonance-Guided Radiotherapy

Recently, hybrid devices have been developed that combine MRI scanners with a LINAC for radiotherapy delivery, bringing to the clinics a novel technique: magnetic resonance-guided radiotherapy. One of such hybrid devices, shown on fig:linac: The ViewRay (Mountain View, CA, USA) MRIdian LINAC system combines a 0.35 T MRI scanner with compact LINAC. Technical details of the system are described by Klüter (2019).

Daily MR imaging provides excellent soft tissue contrast for patient adjustment and also allows for on-table customisation of treatment plans, which is fully integrated into the system's treatment workflow. Automatic beam control during treatment is facilitated by cine MR imaging

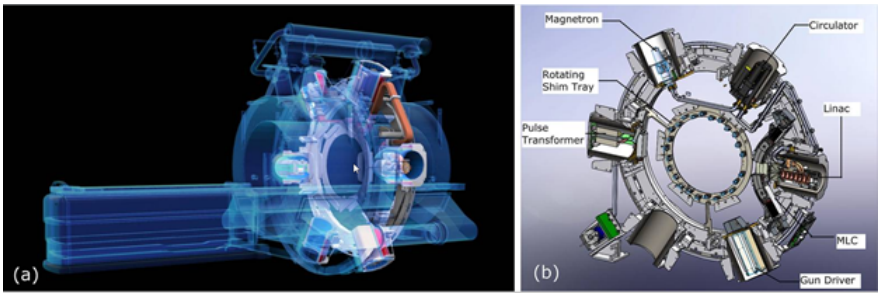


Figure 2.4: (a) Schema of the system with the main hardware components: superconducting double-donut magnet, circular irradiation gantry and patient couch; (b) schema of the irradiation gantry with LINAC components and MLC. The fraction of alligned protons depends on the strenght of the magnetic field, while 0.35 the mmore protons aligned the image quality improves. If we have strong magnetic field, MR takes few minutes due to the low magnetic field. Source: Klüter (2019)

and structure tracking. The new hybrid MR-LINAC system offers superior 3D anatomical imaging for precise tumour delineation and the instantaneous detection of interfractional changes. At the same time, it provides information through cine MRI, enabling continuous monitoring of tumour volume and nearby critical structures throughout the treatment session. Compared to conventional radiotherapy techniques, safety margins and thus irradiated volume can be reduced, lowering the risk of toxicity.

As described earlier, CT images are still a requirement for the treatment planning process, supported by an overlay of an MR image of a patient. The CT image is required to obtain a map of electron density and to calculate the dose distribution in the patient, while the MR image provides better soft tissue contrast that allows precise delineation of tumour and organ contours for treatment planning. During treatment planning, both CT and MR images are taken during the treatment planning visit (Day 0) and co-registered further. The time interval between the two scans is kept as limited as possible (<30 min). The same setup and fixation is used for the patient for both image modalities.

To calculate the radiation dose, the current clinical workflow utilises the CT acquired at Day 0 and produce synthetic deformable CT (dCT) images of the day, based on its co-registration to the most recent MR image, acquired at each treatment visits. Electron density values are assigned to the voxels of the patient's MR images by warping the electron densities of the dCT images. However, the uncertainty of image registration increases with the variation of a patient's anatomy and especially with the presence of tumours. The error in deformable registration can be as high as 3.7 mm, as shown in a study evaluating deformable registration methods by Nyholm et al. (2009) A schematic overview of the current clinical practice of MRgRT at the USZ is shown in Figure 2.5, where the MRIdian LINAC system is in clinical use since 2019.

2.4 Dose Volume Calculation and Evaluation of Dose Distributions

Radiotherapy aims to destroy cancer cells that are dividing. Nevertheless, it also affects the dividing cells of healthy tissue and this damage to the healthy cells causes undesirable side effects. Radiotherapy is about striking a balance between destroying cancer cells and minimising damage

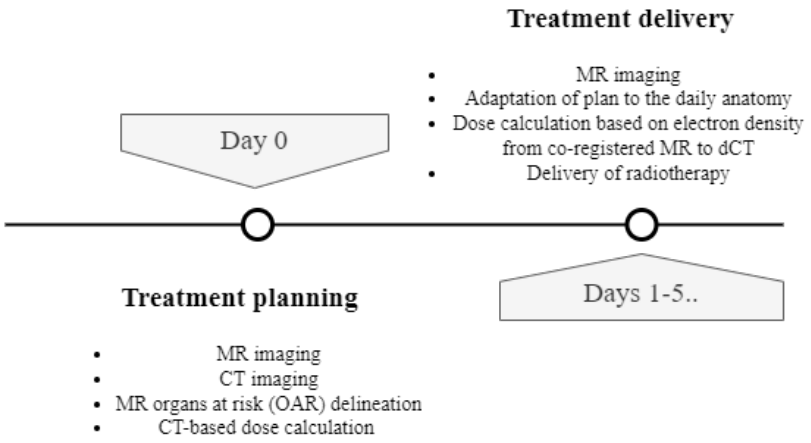


Figure 2.5: Schematic representation of the USZ cancer treatment routine

to healthy tissue, especially organs at risk that are close to the tumour and may be in the irradiation path. An ideal dose distribution scenario exists when the prescribed amount of radiation is delivered to the planning target volume (PTV), which consists of the tumour volume and the margin accounting for organ movement (Antolak and Rosen, 1999), and no healthy tissue is damaged. However, this is not achievable because healthy tissues lying along the irradiation path. The aim of dose optimisation is therefore to get target dose to the PTV and try to deliver as little dose as possible outside the tumour volume.

While planning, the physician defines the target region for the prescription dose and the maximum doses for other structures, by optimising different beam parameters and locating the isocenter of a tumour. The gray (Gy) is used as a unit of the radiation quantity absorbed dose that measures the energy deposited by ionising radiation in a unit mass of matter being irradiated, and is used for measuring the delivered dose of ionising radiation in radiotherapy:

$$1Gy = 1\frac{J}{kg} \tag{2.3}$$

The dose absorbed by a volume depends on it is mass and the amount of energy delivered to the volume. The energy, reaching the volume depends on the attenuation μ that the radiation meets on its way there. The gold standard for dose calculation is the physics-based Monte Carlo method, in which the movements of millions of radiation particles are simulated to produce an estimate of the dose distribution based on the selected beam parameters (Wyckoff et al., 1976; Library, 2011).

Concerning the task of sCT generation, this means that not only the density of a tumour on generated images plays a role, but also the density of tissues on the beam trajectory as well as the densities of OAR volumes. This fact illuminates the importance of assessing dosimetry accuracy while considering the quality of sCT generation. In this regard, per-pixel error metrics could be used to assess the quality of an image as a whole, but this takes into account the errors outside the region of interest, which could be neglected in real-world environments. In addition, deviations in bone tissue that lie outside the beam direction could lead to significant errors in accuracy per pixel due to the high HU values of bone on CT images and their large gradient difference from soft tissue. This fact is particularly important in the abdomen because the ribs contain low hydrogen concentrations and are thus almost invisible on MR scans, adversely affecting the robustness of

the CT algorithm. Hence, only evaluation of dose distributions would provide the most accurate assessment of the potential for integrating the methods into current clinical practice.

Evaluation of the dose distribution for MRgRT treatment plans is performed with the use of 2D isodose line plots over the region of interest on CT images, i.e. PTV and OAR. Several tools have been developed, of which the dose-volume histogram (DVH) is generally considered among the most beneficial (Drzymala et al., 1991). A DVH is a graph that shows the relationship between the volume of an organ or PTV and the dose that the volume is receiving.

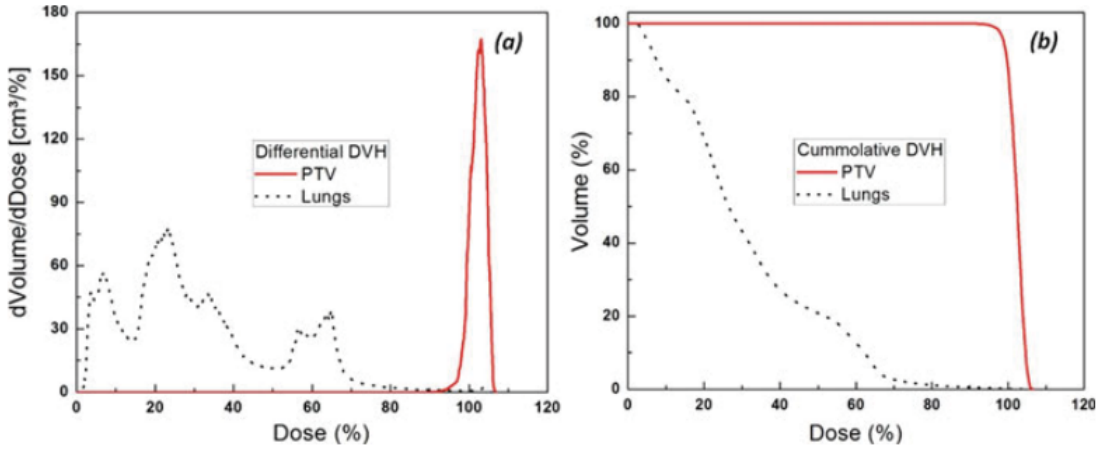


Figure 2.6: Schematics of (a) differential DVHs and (b) cumulative DVHs for the PTV and lungs in a phase-2 mediastinum treatment plan. Source: Hussain and Muhammad (2017a)

There are two options for plotting the dose-volume histogram: differential and cumulative. A differential DVH is a representation of the frequency of occurrence of different dose values. The entire irradiated volume is divided into a finite number of small volume elements (voxels). All voxels that receive a dose from a specific dose range are counted. This process is repeated for all voxels in the volume of interest. A plot showing the number of voxels in each bin and the dose range of the bin is a differential DVH. A cumulative DVH describes how much of the volume receives a dose equal to or greater than the dose in question. Mathematically, a cumulative DVH can be calculated by integrating the differential DVH. Hussain and Muhammad (2017b) provide a neat illustration of the corresponding cumulative and differential DVHs (see Figure 2.6). Radiation oncologists base their decisions for plan approval based on a cumulative DVH as well as a stratified evaluation of 2D dose distributions.

There are a variety of parameters to quantitatively compare the dose distribution calculated with real dCT and generated synthetic CT. Based on the statistical analysis, the spread or dispersion of a dose within a volume could be measured by the following parameters, which are used in the study: D_2, D_{mean}, D_{95} and D_{98} for PTV and OAR. D_2 and D_{98} represents the dose to the 2% and 98% of the volume, respectively. In other words D_{98} indicates that 98% of the target volume receives this dose or higher. In this definition, D_{98} and D_2 are considered as near-minimum and near-maximum doses respectively. While D_{mean} is a mean dose, delivered to a volume. One of the most essential metrics is D_{95}. D_{95} for a PTV indicates that 95% of the tumour volume receives this dose. That means that D_{95} should be as close as possible to the prescribed irradiation dose (Kim and Suh, 2007). Moreover, the dose constraints for the critical structures in abdominal

plans are normally estimated for the set of OAR: stomach, duodenum, bowel, spinal cord (D_2) and liver (D_{98}). The listed above dose distribution parameters provide coverage of the most pertinent components of the DVH-based dosimetric accuracy estimation and enable a matter-of-fact comparison of the sCTs generated by different methods.

Chapter 3

Related work

3.1 Generative Adversarial Networks (GAN)

Medical image synthesis can be formulated as an image-to-image translation task, where a model maps the input image (A) to a target image (B). Among all possible strategies, deep learning methods have drastically improved the current MR-to-CT atlas-based registration employed in clinical practice (Yi et al., 2019). Many of the recently applied DL architectures for image-to-image translation tasks are based on the Generative Adversarial Network (GAN) concept proposed by Goodfellow et al. (2014)

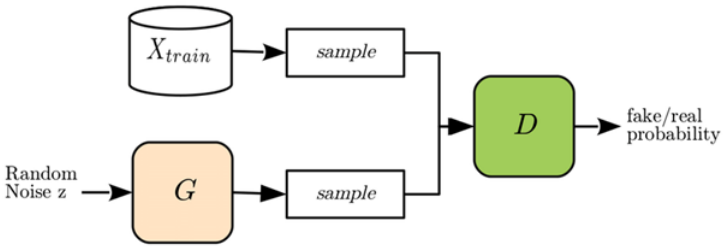


Figure 3.1: Generative Adversarial Network (GAN) concept. Source: Kim (2018)

GANs are generative models designed for direct sampling from the desired data distribution without the need to explicitly model the underlying probability density function. They are composed of two neural networks: the generator G and the discriminator D. The input of G, z, is a random noise sampled from a prior distribution $p_z(z)$, which for simplicity is usually chosen to be a Gaussian or uniform distribution. The output of G is expected to have a visual similarity to the real sample x drawn from the real data distribution $p_{data}(x)$, E is the expectation value. Figure 3.1 shows the interaction between the generator and the discriminator. The training procedure for G is to maximise the probability that D makes an error. This framework corresponds to a minimax two-player game:

$$\min_G \max_D V(D,G) = E_{x\sim p_{data}(x)}[\log D(x)] + E_{z\sim p_z(z)}[\log(1 - D(G(z)))] \tag{3.1}$$

Generative adversarial networks circumvent the difficulty of approximating many hard-to-perform probabilistic computations. In GANs, only backpropagation is used to obtain gradients, no inference is required during learning, and a variety of factors and interactions can be easily

incorporated into the model. The stochastic minibatch gradient descent training of generative adversarial networks is described in Algorithm 1. Source: Goodfellow et al. (2014)

Algorithm 1 Minibatch stochastic gradient descent training of generative adversarial nets. The number of steps to apply to the discriminator, k, is a hyperparameter. We used $k = 1$, the least expensive option, in our experiments.

for number of training iterations **do**

for k steps **do**

- Sample minibatch of m noise samples $\{\boldsymbol{z}^{(1)}, \ldots, \boldsymbol{z}^{(m)}\}$ from noise prior $p_g(\boldsymbol{z})$.
- Sample minibatch of m examples $\{\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(m)}\}$ from data generating distribution $p_{\text{data}}(\boldsymbol{x})$.
- Update the discriminator by ascending its stochastic gradient:

$$\nabla_{\theta_d} \frac{1}{m} \sum_{i=1}^{m} \left[\log D\left(\boldsymbol{x}^{(i)}\right) + \log\left(1 - D\left(G\left(\boldsymbol{z}^{(i)}\right)\right)\right) \right].$$

end for

- Sample minibatch of m noise samples $\{\boldsymbol{z}^{(1)}, \ldots, \boldsymbol{z}^{(m)}\}$ from noise prior $p_g(\boldsymbol{z})$.
- Update the generator by descending its stochastic gradient:

$$\nabla_{\theta_g} \frac{1}{m} \sum_{i=1}^{m} \log\left(1 - D\left(G\left(\boldsymbol{z}^{(i)}\right)\right)\right).$$

end for

The gradient-based updates can use any standard gradient-based learning rule. We used momentum in our experiments.

If the discriminator is trained to optimality before each generator parameter update, then minimising the value function amounts to minimising the Jensen-Shannon divergence between the real data distribution $p_{data}(x)$ and prior distribution $p_z(z)$. This, however, often leads to the vanishing gradient problem when the generator fails due to the fact that an optimal discriminator does not provide enough information for the generator to advances.To solve the well-known vanishing gradient problem in GANs, a number of other learning objectives have been proposed, the most commonly used of which are Least Squares GAN (LSGAN) and Wasserstein GAN (WSGAN) (Hunter, 2018). Where the original GAN uses a log loss, the LSGAN uses an L2 loss (which equates to minimising the Pearson X^2 divergence) while training the discriminator, which forces generator to produce samples toward decision boundary. In the LSGAN, the loss for real samples should be lower than the loss for fake samples. This allows the LSGAN to target fake samples that have a really high margin. The authors of the LSGAN claims (Mao et al., 2017) that the LSGANs are able to generate higher quality images than original GANs and solve the vanishing gradient problem. Another learning objective (Wasserstein GAN) was introduced to solve the mode collapse problem of original GANs. Mode collapse occurs when the discriminator gets stalled in a local minimum and fails to find the optimal strategy, while each iteration of the generator is overoptimised for a particular discriminator and the discriminator never manages to finagle its way out of the local minimum. As a result, the generators rotate through a small set of output types. The Wasserstein loss alleviates mode collapse by training the discriminator to optimality without worrying about vanishing gradients (Arjovsky et al., 2017). It leverages the Wasserstein distance to produce a value function which has better theoretical properties than the original GAN:

$$\min_G \max_{D \in D_s} L(D, G) = E_{x \sim p_{data}(x)}[D(x)] - E_{z \sim p_z(z)}[\log(1 - D(G(z)))] \tag{3.2}$$

where D_s is the set of 1-Lipschitz functions. In that case, under an optimal discriminator, minimis-

ing the value function with respect to the generator parameters minimises Wasserstein distance between the real data distribution $p_{data}(x)$ and prior distribution $p_z(z)$. WGAN requires that the discriminator (referred to as the critic in that paper) must lie in the space of 1-Lipschitz functions, which the authors enforce via weight clipping, although this can lead to gradient extinction or gradient explosion problems (Qin and Jiang, 2018). Another research called "Improved WGAN" proposes instead of the weight clipping but rather add a penalisation term to the norm of the gradient of the critic function(Gulrajani et al., 2017) to achieve Lipschitz continuity. To circumvent tractability issues, we enforce a soft version of the constraint with a penalty on the gradient norm for random samples $\widehat{x} \sim p_{\widehat{x}}$, where $\widehat{x}$ represents a soft version of the constraint with a penalty on the gradient norm for random samples and $p_{\widehat{x}}$ is a uniform sampling distribution along straight lines between pairs of points sampled from real data distribution $p_{data}(x)$ and prior distribution $p_z(z)$. The objection function of Wasserstein GAN-GP is:

$$\min_G \max_{D \in D_s} L(D,G) = E_{x \sim p_{data}(x)}[D(x)] - E_{z \sim p_z(z)}[\log(1 - D(G(z)))] + \lambda E_{\widehat{x} \sim p_{\widehat{x}}} \left[\left(\|\nabla_z D\left(\widehat{x}\right)\|_2 - 1 \right)^2 \right] \tag{3.3}$$

There are several other research works focusing on the theoretical improvement of GANs to overcome the well-known GAN stability training problems, which are further embedded in the numerous architectures and utilised across wide range of domains (Hunter, 2018). Regarding medical image analysis, various applications of GANs have been developed for image preprocessing, organ segmentation, anomaly detection, domain adaptation and translation of an image of one modality to an image of another modality, including the synthesis of sCT, as it is demonstrated in Figure A.3. The current state of DL-based architectures for generating sCT includes pix2pix: a conditional generative adversarial image-to-image network that requires paired images from two modalities co-registered with voxel-wise correspondence (Isola et al., 2017). The main motivation of researchers has long been to overcome the main limitation: the lack of aligned image pairs for domain translation, which is particularly peculiar to medical images. An alternative architecture that accommodates the additional cycle consistency loss has been developed more recently and is called Cycle Consistent Generative Adversarial Network or CycleGAN (Zhu et al., 2017). It requires only images from each modality rather than almost perfectly aligned paired images of the same patient. In the next sections, the architectures and the differences in loss function objectives for the purpose of sCT generation are discussed, as well as the new patch-bathed method for unpaired image-to-image translation, which is known as CUT (Park et al., 2020).

3.2 Image-to-Image Conditional Generative Adversarial Network (pix2pix)

GANs are generative models that learn a mapping from random noise vector z to output image y, $G : z \rightarrow y$. In contrast, conditional GANs (Li et al., 2020) learn a mapping from observed image x and random noise vector z, to $y, G : x, z \rightarrow y$.

The pix2pix model is a conditional image-to-image generative adversarial network that requires paired images from two modalities that are co-registered with voxel-wise correspondence (Isola et al., 2017). In addition to the adversarial GAN losses consisting of the generator loss and the discriminator loss (real versus fake image pairs), it includes an additional loss based on the

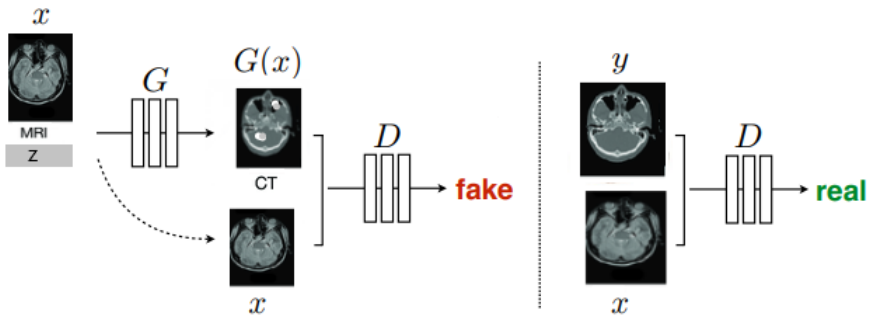


Figure 3.2: Training a conditional GAN. The discriminator, D, learns to classify between fake (synthesised by the generator) and real tuples. The generator, G, learns to fool the discriminator. Unlike an unconditional GAN, both the generator and discriminator observe the input MRI. Inspired by original pix2pix paper by Isola et al. (2017)

absolute difference between the generated image and the original paired image ($L1$ norm loss). The $L1$ norm loss is expressed as:

$$L_{L1}(G_S^T) = E_{x,y}(|y - G_S^T(x)|_1) \tag{3.4}$$

where G_S^T is the generator network that produces images corresponding to the target modality from the source modality images, and E is the expectation that depends on both x, the set of source modality images, and y, the set of target modality images. The source modality images x in our tasks are MR images and the target modality images y are CT images. The adversarial loss penalises at the scale of subimage patches and is expressed as:

$$L_{GAN}(G_S^T, D_T) = E_{x,y}(\log D_T(x, y)) + E_x \log(1 - D_T(x, G_S^T(x))) \tag{3.5}$$

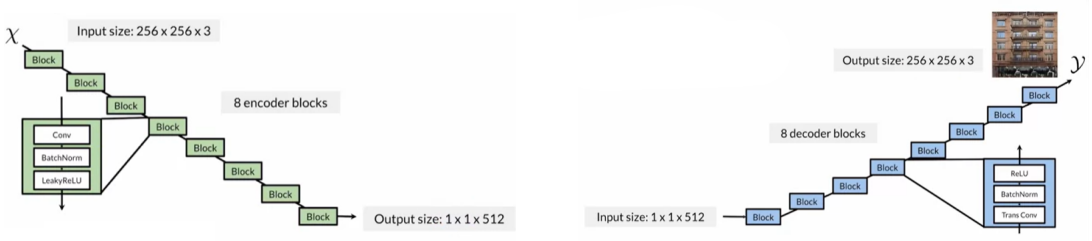
where D_T is the target modality discriminator that aims to distinguish between real and fake images, E is the expectation value. When using pix2pix networks, the number of samples in the source and target domains must be the same since the data sets must be aligned with each other. The adversarial loss is calculated using the binary cross entropy cost function. The final cost function, which is used to optimise the network, is a weighted summation of the aforementioned losses:

$$\Theta_{G,D} = \arg\min_G \max_D (L_{GAN}(G_S^T, D_T) + \lambda L_{L1}(G_S^T)) \tag{3.6}$$

where λ is the user-defined weighting factor for the $L1$ loss (Klages et al., 2020).

The basic pix2pix implementation consists of a U-Net-based generator and a PatchGAN-based discriminator. One of the reasons pix2pix first achieved task-agnostic image translation supporting multiple image-to-image translation tasks, including biomedical, is an architecture of the generator.

U-Net-based generator. U-Net is a convolutional neural network that was initially developed for segmenting biomedical images by Ronneberger et al. (2015). It is one of the most cited architectures in medical image analysis for a range of applications. The special feature of the U-net that has a positive impact on the accuracy of the models is the introduction of the skip connector in the standard encoder-decoder architecture, which allows the information from the earlier layers to be transferred to later layers, while the upsampling process in the decoder is done by concatenating current layer with the respective encoder layer resolution. In addition, skip connectors improve



(a) Encoder (b) Decoder

Figure 3.3: PIX2PIX GENERATOR. U-Net-based generator architecture including encoder (a) and decoder (b) parts. Source: Sharon and Eda Zhou (2020)

gradient flow and help to solve the vanishing gradient problem. In each block of the pix2pix encoder, the spatial information is reduced by a factor of 2. The output size at the end is 1*1*512. Each of the blocks contains the convolution layer (stride=2), batch normalisation and LeakyReLU as activation. The decoder part consists of the same number of blocks as the encoder to be able to rebuild the image in the same size. Each block of the decoder consists of the transposed convolution, which performs a corresponding upsampling procedure. The transposed convolution is followed by a BatchNorm and subsequently by a ReLU activation function. Batch normalisation has been shown to be essential to train both networks avoiding mode collapse (Radford et al., 2015). Dropout is added to the first three blocks of the decoder. Dropout randomly deactivates different neurons at each iteration of training so that different neurons can learn in a stochastic manner. In the end, either the real sample or the generated sample is concatenated with what was inputted into the generator as an input condition and passed on to the discriminator.

PatchGAN-based discriminator. A PatchGAN discriminator penalises the structure only at the patch level. Such a discriminator effectively models the image as a Markov random field, assuming independence between pixels separated by more than one patch diameter. This discriminator attempts to classify whether each N × N patch within an image is real or fake. The discriminator patches are applied to the image by convolution, combined to form the matrix of responses, and all responses are further averaged to obtain the final result D. The advantages of the discriminator include speed due to the smaller size of the patches and the ability to control the quality of the large images.

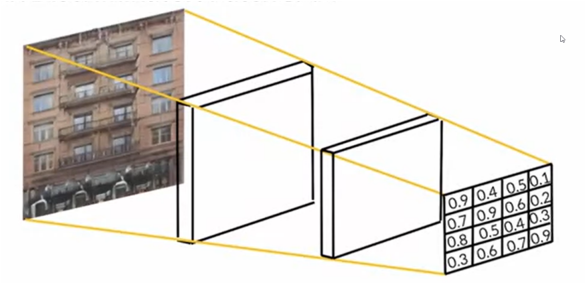


Figure 3.4: Schematic representation of the PatchGAN. Source: Sharon and Eda Zhou (2020)

One of the major difficulties when applying pix2pix to medical applications is that it requires co-registered image pairs. Another architecture, CycleGAN, has the potential to leapfrog the

limit of using a perfectly aligned dataset for training and could achieve similar performance on an image and dose basis for medical image-to-image translation tasks (Largent et al., 2019).

3.3 Cycle-Consistent Generative Adversarial Network (CycleGAN)

The CycleGAN model is a cycle-consistent generative adversarial network, which was introduced by Zhu et al. (2017). CycleGAN is similar to pix2pix in that it uses the same foundational network blocks, but it requires only images from each modality, rather than nearly perfectly aligned paired images from the same patient. CycleGAN seeks to learn not the specific transformation of each individual pixel, but the transformation of the image properties as a whole. The requirement for alignment of images is eliminated by an additional cycle consistency loss, so that each image that passes through the pair of generators attempts to reproduce itself.

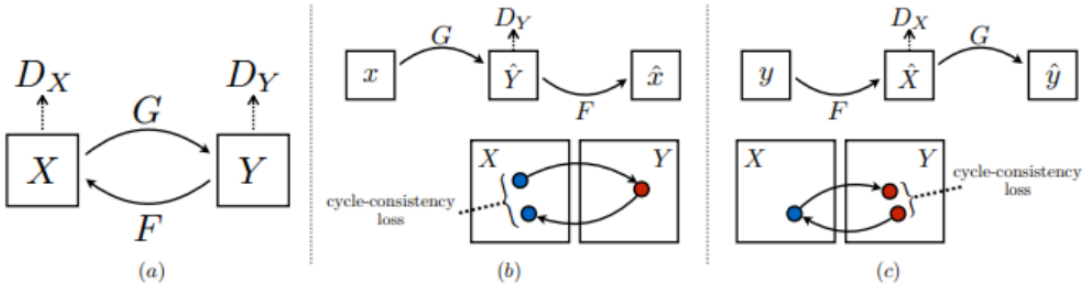


Figure 3.5: Schematic representation of the cycle consistensy loss: given a real image x in X, if the two generators G and F are good, mapping it to domain Y and then back to X should give back the original image x, i.e., x → G(x) → F(G(x)) ∼ x. Similarly, the backward direction should also have y → F(y) → G(F(y)) ∼ y. Source: Wang and Lin (2018)

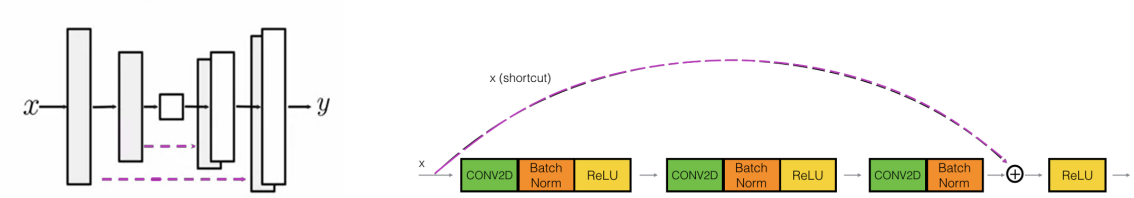
Cycle consistency in both directions is required for this training. The sequence is for the sCT generation task as follows: MR → sCT1 → sMR1 and CT → sMR2 → sCT2. Cycle consistency is ensured by minimising the $L1$ norm losses between the output synthesised images (sMR1, sCT2) and the corresponding input modality images (MR and CT, respectively). The result is a single cycle network consisting of a pair of GANs operating on the two image modalities. The cycle consistency loss is expressed as:

$$L_{cyc}(G_x^y, G_y^x) = E_x(|x - G_y^x(G_x^y(x))|_1) + E_y(|y - G_x^y(G_y^x(y))|_1) \tag{3.7}$$

where G_x^y is the generator in the direction of MR → CT as x is an input modality (MR) and y is output or target modality (CT). Accordingly, G_y^x depicts the generator in the direction CT → MR. The GAN loss is expressed as:

$$\begin{aligned} L_{GAN}(G_x^y, G_y^x, D_x, D_y) = E_y[\log(D_y(y))] + E_x[\log(1 - D_y(G_x^y(x)))] \\ + E_x[\log(D_x(x))] + E_y[\log(1 - D_x(G_y^x(y)))] \end{aligned} \tag{3.8}$$

where D_x is the input modality discriminator (MR) and D_y is the target modality discriminator (CT). An additional identity loss that forces the intermediate images to have similar intensities to



(a) U-Net skip connection via concatenation

(b) ResNet skip connection via addition

Figure 3.6: SKIP CONNECTIONS. The violet arrows show the differences in skip connections between the U-net architecture (a) and the ResNet (b)

the actual intermediate group is expressed as follows:

$$L_{identity}(G_x^y, G_y^x) = E_y(|y - G_x^y(y)|_1) + E_x(|x - G_y^x(x)|_1) \tag{3.9}$$

The final cost function, which is used to optimise the network, is a summation of the aforementioned losses:

$$\Theta_{G,D} = \arg\min_G \max_D L_{GAN}(G_x^y, G_y^x, D_x, D_y) + \lambda_{cyc} L_{cyc}(G_x^y, G_y^x) + \lambda_{identity} L_{identity}(G_x^y, G_y^x) \tag{3.10}$$

where λ_{cyc} and $\lambda_{identity}$ are the user-defined weighting factors. These cyclic constraints are less stringent than the voxel-wise constraints of the pix2pix model, however the objective of creating images in the second modality based on the input images remains. (Klages et al., 2020).

By devoting attention to the architecture of specific network components, the generators could be implemented using the U-Net described in Section 3.2 or as in the implementation by Zhu et al.[1] using a ResNet-based generator (see Figure A.4). ResNet-based generator consists of multiple ResNet blocks between a series of basic downsampling/upsampling blocks in an encoder-decoder approach (see Figure 3.3). Instead of the U-Net skip connection aimed to recover spatial information lost during downsampling, where the parts of the downsampling operations are concatenated with the corresponding upsampling operations, in ResNet the skip connection aimed to preserve features learned in earlier layers. That is why it skips every two consecutive convolutions within the downsampling or upsampling operations by itself and "flatten" the model therefore. Detailed representation of the ResNet9 building blocks with a skip connection is shown in Figure 3.6b. This aids in solving the vanishing gradient problem by providing an alternative path for the gradient to flow through (He et al., 2016).

The discriminator of CycleGAN is a convolutional PatchGAN classifier, similar to pix2pix, models high frequency image structure in local patches and only penalises structure at the scale of image patches (see Figure 3.4). It consists of 3 convolutional layers with different filter sizes but the same kernel sizes and strides, followed by fully connected layers. LeakyReLU is used as the activation function for non-linearity and a batch normalisation layer for the convolutional layers, sigmoid activation function is used in the last fully connected layer.

One of the important notes on the training strategy of CycleGANs, implemented by the Zhu et al.[1], is that they are using the pool aspect, introduced by Shrivastava et al. (2017). They keep an image buffer that stores the 50 previously generated images and increase the discriminators by using a history of the images generated instead of using the ones produced in latest generators cycle. This helps to reduce the oscillations of the loss over time and thus avoid model collapse.

[1]https://github.com/junyanz/pytorch-CycleGAN-and-pix2pix

This problem occurs when the generator learns to map several different input values to the same output (Goodfellow, 2016).

The CycleGAN demonstrates comparable results to models that require perfectly aligned image pairs, which has the potential to eliminate the additional cost of co-registering images associated with the task of sCT generation. This is where one of the main advantages of such an architecture becomes apparent. It performs well with image style changes, however, it has difficulties with significant geometric changes due to the cyclic architecture and domain invariance. The presence of two cycles is the underlying reason for the longer training time.

3.4 Contrastive Learning for Unpaired Image-to-Image Translation (CUT)

Contrastive Unpaired Translation (CUT) utilises a contrastive learning-based framework that aims to associate input and output fields, whereby 'query' refers to an output patch and 'positive' and 'negative' are corresponding and non-corresponding input patches, sampled within the image (see Figure 3.7). CUT framework was introduced by Park et al. (2020) on Computer Vision and Pattern Recognition conference.

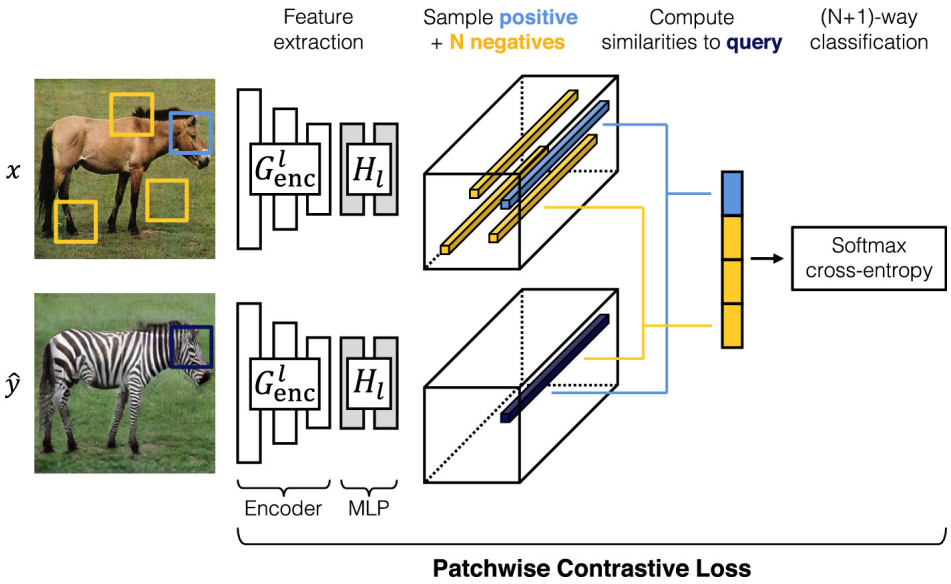


Figure 3.7: The schematic representation of the CUT architecture, utilising novice patchwise contrastive loss. Source: Park et al. (2020)

In CUT, the generator is divided into two parts, where the encoder G_{enc} learns to capture domain-invariant concepts, and the decoder G_{dec} learns to generate domain-specific patterns. These are applied in sequence to generate an output image $y = G(z) = G_{dec}(G_{enc}(x))$. The adversarial loss is employed to encourage the output to be visually similar to images from the target domain:

$$L_{\text{GAN}}(G, D, X, Y) = E_{\boldsymbol{y}\sim Y} \log D(\boldsymbol{y}) + E_{\boldsymbol{x}\sim X} \log(1 - D(G(\boldsymbol{x}))) \tag{3.11}$$

The authors propose using the first half of the CycleGAN generator as G_{enc}. One of the key novelties of this approach is the use of a noise contrastive estimation framework to maximise the mutual information between the input and output images of the generator. This is done by sampling a patch from the image produced by generator (query patch) and comparing it to the input patch at the same location (positive patch). In addition, N patches are sampled from other locations in the input image, which are referred to as "negative". By sampling negative patches within the image, the authors hypothesise that the encoder does not need to model large intra-class variation. They then use G_{enc} together with a two-layer MLP to encode both input and output fields in a common embedding space. The query, positive, and N negatives in generator are mapped to K-dimensional vectors $\boldsymbol{v}, v^+ \in R^K$ and $v^- \in R^{N \times K}$, respectively. ${v^-}_n \in R^K$ denotes the n-th negative. They normalise vectors onto a unit sphere to prevent the space from collapsing or expanding. An (N+1)–way classification problem is set up, where the distances between the query and other examples are scaled by $\tau = 0.07$:

$$\ell(\boldsymbol{v}, v^+, v^-) = -\log \left[\frac{\exp(\boldsymbol{v} \cdot v^+/\tau)}{\exp(\boldsymbol{v} \cdot v^+/\tau) + \sum_{n=1}^{N} \exp(\boldsymbol{v} \cdot {v^-}_n/\tau)} \right] \tag{3.12}$$

It is fed as logits to a cross-entropy loss function that represents the probability of selecting a positive over a negative. The PatchNCE loss uses the matching of corresponding input-output patches in the generator at a given location:

$$L_{\text{PatchNCE}}(G, H, X) = E_{\boldsymbol{x} \sim X} \sum_{l=1}^{L} \sum_{s=1}^{S_l} \ell(\hat{\boldsymbol{z}}_l^s, \boldsymbol{z}_l^s, \boldsymbol{z}_l^{S \setminus s}). \tag{3.13}$$

Beyond that, a multi-layered approach is proposed, in which not only image patches from the input or output image in the generator are sampled, but also from deeper layers. The external NCE loss uses image patches that differ from the rest of the data set ("negative" patches):

$$L_{\text{external}}(G, H, X) = E_{\boldsymbol{x} \sim X, \tilde{\boldsymbol{z}} \sim Z^-} \sum_{l=1}^{L} \sum_{s=1}^{S_l} \ell(\hat{\boldsymbol{z}}_l^s, \boldsymbol{z}_l^s, \tilde{\boldsymbol{z}}_l), \tag{3.14}$$

where dataset negatives $\tilde{\boldsymbol{z}}_l$ are sampled from an external dictionary Z^- from the source domain. Detailed description could be found in (He et al., 2020).

Additionally, PatchNCE loss $L_{\text{PatchNCE}}(G, H, Y)$ utilised on images from domain Y to prevent the generator from making unnecessary changes. This loss is essentially a learnable, domain-specific version of the identity loss, commonly used by previous unpaired translation methods (Taigman et al., 2016). The final cost function, which is used to optimise the CUT network, is a summation of the aforementioned losses:

$$\mathcal{L}_{\text{GAN}}(G, D, X, Y) + \lambda_X \mathcal{L}_{\text{PatchNCE}}(G, H, X) + \lambda_Y \mathcal{L}_{\text{PatchNCE}}(G, H, Y). \tag{3.15}$$

As for the specific architecture implementation of the blocks, it is the same as the setting of CycleGAN with ResNet-based generator and PatchGAN described in the previous section, except that the cycle consistency loss is replaced by the contrastive loss. Compared to CycleGAN, CUT allows for one-sided translation in the unpaired image-to-image translation setting, while improving quality and reducing training time. Besides, one-way translation could be advantageous for medical image-to-image translation, as tissue features are usually consistent for each patient (except for tumours), while there is high variability in physiology, which very often affects even co-registered image pairs of different modalities and may complicate cycle-consistent training.

3.5 Application of GANs for the sCT Generation Tasks

The desired elimination of computed tomography from the MRgRT pipeline with the aim of reducing the irradiation of healthy tissues has been a challenge for researchers for a considerable time. Currently, there are 3 basic approaches to estimating electron density maps, which are required for tumour radiation dose calculation:

- Bulk density approach, which is based on manual segmentation and further assignment of a homogeneous, predefined density to each region. The advantage of this approach is the possibility of quality control by experts. However, the quality of the density assignment and thus of the calculated dose estimate depends entirely on the quality of the segmentation. This means that incorrect assignment at the margins of the tissue would affect the dose distribution in overall terms. The bulk approach requires MR images with a special sequence: the ultra-short echo time (UTE), which is characterised by a long acquisition time and negative effects on the quality of the bone representation on the images. Finally, it is tedious for human experts (Kang et al., 2017).

- Multi-atlas based approach, which involves rigid and nonrigid mapping of atlas CT images onto a target MR image. To accurately map intricate anatomy, an atlas dataset is created from co-registered CT-MR image pairs. An initial step involves pairwise mapping of MR atlas images to the target MR image. The advantages of this approach include its applicability to the entire image population along with very reasonable quality. This approach does not require the daunting manual delineation efforts by experts. That is why this approach is widely used in hospitals for radiation dose correction during patient treatment. The drawbacks include the need for co-registered image pairs, the mathematical complexity of deformable registration and quality errors for some of the patients who differ substantially from the "average atlas representative" of the population (Pham et al., 2000).

- Machine learning based approach where algorithms learn the intensity mapping of MR images to electron density maps usually through highly nonlinear systems. Neural networks show the high capability to estimate the electron density maps. One of the advantages is the fully automatic training process. Though, the best results have been obtained with the algorithms that require co-registration of the image modalities so far (Klages et al., 2020)

The existing studies analyse the results of substitute CT generation across different methods between the bulk density and multi-atlas based approaches (Edmund and Nyholm, 2017; Johnstone et al., 2018) or between multi-atlas based approaches and machine learning based approaches (Han, 2017; Arabi et al., 2018). No studies were found investigating the difference in the results across 3 methods for the abdomen area.

Farjam et al. (2019) provided a comparison of an atlas-based sCT generation and a bulk density approaches for the pelvic region. The bulk density-based approach was reported to produce very sharp and neat images, while the proposed atlas-based approach produced a rather blurred image. The largest discrepancies were found in the bony structures. The proposed multiatlas approach outperforms the bulk density-based approach in terms of Hounsfield unit (HU) assignment and performs slightly better in terms of reproducing the dose distribution from the original plan (see Table 3.1).

Han (2017) were investigating the difference in performance of the atlas-based methods and machine learning U-Net architecture in brain images. They demonstrated that NN-based methods were able to surpass others and produce acceptable results, with mean absolute errors (MAE) of 84.8 HU compared to 94.5 HU for deep convolution neural network (DCNN) and Atlas-based methods correspondingly Fard et al. (2021). A year later, Arabi et al. (2018) tested the same model with different atlas-based methods for a bigger cohort of the patients (n=38) and confirmed that

	Bulk density-based approach	**Atlas-based approach**
over entire CT	65 ± 5	47 ± 5
bones	172 ± 9	116 ± 12
fat	43 ± 7	36 ± 6
muscles	47 ± 5	42 ± 4

Table 3.1: The average of mean absolute errors (MAE,mean ± SD) in HU between the atlas-based and bulk density based approach for MR-only radiotherapy of pelvis anatomy, reported by Reza Farjam et al. Farjam et al. (2019)

	Atlas-based approach	**DCNN-based**
body	42.4 ± 8.1	32.7 ± 7.9
bones	130.2 ± 23.4	119.9 ± 22.6

Table 3.2: The average of mean absolute errors (MAE,mean ± SD) in HU between the atlas-based (ALWV-Iter Burgos (2017))method and machine learning-based method (U-Net architecture) for MR-only radiotherapy of pelvis anatomy, reported by Arabi et al. (2018)

the deep convolution neural network in most of the cases outperform the multi-atlas based methods (see Table 3.2). However, the biggest difference is still in the bone region. This could be explained by the low hydrogen content of the bones and the low intensities on the MR images, which closely resemble air.

In 2016, one of the first research groups were applying GANs for the purpose of the sCT generation. Nie et al. (2017) developed a pix2pix like architecture, where the generator was consisting of the encoder and decoder parts. Moreover, they addressed the issue of image misalignment by incorporating an image-wise loss together with a voxel-wise loss component. It offers an additional auto-content model, which gives to the classifier additional context information. They trained it on 16 subjects from the brain dataset and 22 subjects from the pelvic dataset. In both cases, GAN outperformed atlas-based approaches, with an average MAE of 171.5 HU and 92.5 HU for atlas-based methods and GAN for the brain dataset and 66.1 and 39.0 for pelvic dataset correspondingly. It is important to note that the errors in the different fields of view are not comparable due to differences in the set of organs imaged, physiological differences and, in addition, differences between device settings.

Wolterink et al. (2017) applied the CycleGAN architecture with the ResNet generator and PatchGAN discriminator for the purposes of the sCT generation from brain MR images. One of their intriguing findings is that training on unpaired data actually shows the better performance when using the CycleGAN architecture: "Qualitative analysis showed that CT images obtained by the model trained with unpaired data looked more realistic, contained less artefacts and contained less blurring than those obtained by the model trained with paired data." Authors hypothesised that errors in the co-registration of the MR and CT pairs could be responsible for the worse performance of aligned methods.

Klages et al. (2020) employed pix2pix and CycleGAN architectures to assess the effects of multiple combination strategies on accuracy for patch-based synthetic computed tomography generation for MRgRT planning in head and neck (HN) cancer patients . Their results showed that pix2pix slightly outperformed CycleGAN with the corresponding MAE of 156.3 HU and 165.2 HU. They suggest that this may be due to the specific loss functions applied (they used L1 losses per pixel). They also tested different input configurations and found that combining the three orthogonal views improved the results compared to training with axial slices only. As preprocessing, they used histogram standardisation techniques and applied intensity clipping.

The assessment of the dosimetric accuracy results revealed that the NNs have a high potential for clinical applicability. In studies by Maspero et al. (2020), the authors found that has accurate MR-based dose calculation using a combination of three orthogonal planes for sCT generation is feasible even for paediatric brain cancer patients (with mean MAE 61.0 HU in body), when training on a heterogeneous dataset with help of pix2pix architecture. The mean prescribed dose difference was withing the acceptable norm.

Several studies have investigated the effects of different loss functions on sCT generation. Hiasa et al. (2018) integrated gradient consistency loss (one of the perceptual losses) into training to improve accuracy at the boundaries for CycleGANs. The results show that gradient consistency loss slightly improves MAE. They investigated CT generation for orthopaedic purposes and applied a rather strong clipping of the CT intensity in the range [-150, 350] HU as one of the preprocessing steps. Following the same idea, Lei et al. (2019) introduced structure-consistency loss (MPD), which extracts structural features from the image defining the loss in the feature space to keep the spatial information. They employs both, conditional GAN and CycleGAN and test the change in the architecture of the generator, changing the ResNet part on the proposed Dense block. Lei et al.'s results showed improvements in this sense relative to other unsupervised methods. GD loss function minimises the difference of the magnitude of the gradient between the synthetic image and the original planning CT. For the proposed method, the mean MAE between sCT and CT were 55.7 Hounsfield units (HU) for 24 brain cancer patients and 50.8 HU for 20 prostate cancer patients. Kang et al. (2021) introduces perceptual loss using some of the ResNet blocks of the discriminator in CycleGAN for his calculations. It presents great results and outperformed the U-net with per-pixel L1 loss in terms of errors and similarities for the purpose of sCT generation, and dose estimation for treatment planning of both, thorax and abdomen. Curiously, the generation of the bone structures has been improved, but there remain certain glitches. Referring to the applications of per-pixel loss functions, where inter-pixel errors are calculated, and perceptual loss functions, where higher level differences such as content and stylistic discrepancies are compared, it is worth pointing out that the task-specific loss functions were implemented too for the sCT generation task. Farjam et al. (2021) introduced task-specific loss functions, based Fuzzy-c-means (FCM) clustering was then utilised to classify voxels into fat, muscle, and bone. During the every learning cycle, the wrong classification of every tissue type was penalised as the training objective. The new loss function calculation improved the MAE by more than 18% in high density areas, compare to L1: MAE (mean ± SD) equals to 29.68 ± 4.41, 16.34 ± 2.67, 23.36 ± 2.85, and 105.90 ± 22.80 HU over the entire body(no air), fat, muscle, and bone tissues correspondingly. Although they used the U-Net CNN architecture and not its GAN version, the results they show deserve further investigation, especially in terms of comparing the performance of CNN and its GAN version. Among the limitations of the approach, it is certainly worth noting that their used manual landmark-based standardisation technique for MRI intensities.

In the largest literature reviews on deep learning methods for CT generation, such as (Spadea et al., 2021; Boulanger et al., 2021; Fard et al., 2021), researchers emphasise the importance of different preprocessing methods and their influence on the final results. The evaluation of 4 MRI standardisation approaches as well as the impact of distortion field correction on sCT generation results was investigated by Andres et al. (2020) on one of the largest cohorts with more than 400 brain volumes . The impact of key parameters was assessed on the final results obtained with deep CNNs. Authors drew a number of conclusions regarding preprocessing. Andres et al. found that correcting the bias field (an undesirable artefact primarily arises from the improper image acquisition process or the specific properties of the imaged object (Song et al., 2017)) did not greatly affect the results for the data they used. They suggest that this may be due to the NN learning how to estimate the bias field of a device. Another finding was that the white-stripe MRI normalisation technique outperformed other methods. This method was presented by Shinohara et al. (2014) as one of the biologically motivated normalisation techniques. The aim of the method

is to minimise the discrepancy between the distributions of intensities across subjects and visits within tissue classes in the brain.

One of the most common methods that can be used in the abdominal area inter-subjects settings, is the Nyul method (Nyúl et al., 2000) showed the second best performance. Not only the lack of specific methods for normalising tissue intensities, but also considerable differences in the shape and position of non-rigid organs and air pockets make the abdominal cavity one of the most difficult regions for training neural networks. Figure 3.9, compiled by Boulanger et al. in 2021, highlights that the abdomen is one of the least studied anatomical locations. One of the best results (MAE = 60.4 HU) for abdominal sCT generation was obtained with a multichannel conditional GAN architecture with ResNET-based generator and L1 loss by Xu et al. (2019). However, they used 4 MR sequences were reconstructed for each subject: fat, water, in-phase (IP), and opposedphase (OP), giving segmentation-like MR inputs. It was possible for them because they trained models on high field MR images.

This study setting does not allow for combining the different MR sequences due to lower quality of the MR images coming from hybrid devices. Coupling LINAC and MR systems was a technical challenge. It required compromises in the image quality and scanning time. For the MRIdian LINAC the solution was to preserve the best quality of the radiation unite and to use a lower magnetic field (0.35T). Advantages of low field MR include reduced costs, better patient access, and greater safety. In this case, high quality examinations can be achieved using appropriate protocols and investing more scanning time than with high-field MR systems. The main disadvantage of low field MR is the reduced signal to noise ratio compared with high-field systems, which could cause additional acquisition artifacts (Hori et al., 2021; Konar and Lang, 2011).

Spadea et al. (2021) showed that the number of studies, investigating DL approaches for sCT generation, based on low field MRI, is 6.5% (0.3-1T), while in all other studies the MR inputs of higher quality were utilized. The best performing study in a similar setting to this study (0.35T MR images as input) was performed by Davide Cusumano et al. Cusumano et al. (2020). They utilised Pix2Pix architecture with U-Net generator and PatchGAN disriminator. Inclusion criteria for each slice were the absence of artefacts, the absence of blurred areas, high correspondence of bony and internal anatomy, as well as air pockets. A total of 9950 images were considered for training the network: 4848 (2681 pelvis, 2167 abdomen) slices were utilised for training the neural network. The body contour, the bones and the soft tissues for the evaluation of the experiment were manually delineated. They achieved the MAE equals to 78 HU within the body contour (see Table 3.3) and DVH parameters Mean Difference <1% (see Table 3.8). One of the best results, which employs CycleGAN architecture in the area of abdomen with 0.35T MR images as input, is reported by Kang et al. in 2021 Kang et al. (2021). Among the most captivating aspects of their study is that they used a mixed dataset of CT and MR images of the pelvis, thorax and abdomen with the aim of preventing the overfitting of networks specialised in a particular area. They used perceptual loss, which examines the discrepancy between high-dimensional representations of images extracted by a CycleGAN discriminator. Moreover, they also utilised three adjacent axial layers as NN input to avoid 3D discontinuity. They achieved the MAE equals to 58.8 HU within the body contour, while norm it by the total number of image voxels (see Table 3.3) and DVH parameters mean difference equals to +/- 0.6% for PTV and equals to +/- 0.15% for OAR (duodenum, stomach, liver).

To the best of our knowledge, this is the first study that employs CUT for MR-based synthetic CT generation in abdominal area.

In conclusion, the deep learning methods described above aim to solve the problem of irradiating healthy tissue in MRgRT by synthetic CT generation based on the MR images. The GANs showed lower geometric errors and higher dosimetric accuracy in sCT generation compared to the bulk density and atlas-based methods currently being used in clinical practice. The CycleGAN architecture, which does not require aligned image pairs, shows comparable performance

	Pix2Pix
MAE body	78.71 ± 18.46
MAE bones	152.71 ± 30.14
MAE soft tissue	53.89 ± 10.7

Table 3.3: The average of mean absolute errors (MAE,mean ± SD) in HU of sCT generation, utilising pix2pix architecture for MR-only radiotherapy of abdomen anatomy, reported by Cusumano et al. (2020)

	CycleGAN
MAE	58.8 ± 4.4
PSNR	26.3 ± 0.7
SSIM	0.91 ± 0.01

Table 3.4: The average of geometrical errors (mean ± SD) in HU of sCT generation in abdomen, utilising CycleGAN architecture for MR-only radiotherapy, reported by Kang et al. (2021)

to pix2pix, which requires co-registered image pairs. Many parameters, such as the NN architecture, different loss functions and the preprocessing methods, could have a positive impact on the results of the image-to-image translation tasks. GANs have experienced the most difficulty with sharp translation of the bone region. The next steps in research on GAN-based sCT generation could be the introduction of a robust NN architecture for the abdominal area and its optimisation.

DVH Indicator	Abdomen	
	Mean Difference (SD)	Range
PTV		
V 95% [%]	−0.28 (1.06)	−2.49; 1.16
D 98% [Gy]	−0.05 (0.23)	−0.39; −0.4
D 50% [Gy]	−0.08 (0.22)	−0.52; 0.4
D 2% [Gy]	−0.13 (0.3)	−0.59; 0.39
Stomach/Rectum		
D 98% [Gy]	0.01 (0.02)	−0.02; 0.05
D 50% [Gy]	−0.04 (0.10)	−0.41; 0.01
D 2% [Gy]	0 (0.13)	−0.38; −0.19
Duodenum/Bowel		
D 98% [Gy]	0.02 (0.03)	0; −0.13
D 50% [Gy]	0.05 (0.16)	−0.02; −0.7
D 2% [Gy]	−0.04 (0.23)	−0.59; −0.36

Figure 3.8: Mean values of volume and dose difference calculated between sCT and CT for all the DVH indicators considered. Absolute dose values were reported in Gy for all the parameters investigated except for V95% of PTV. where the volume percentage difference was considered. For each DVH parameter the standard deviation (SD) and the corresponding range was also reported. Source: Cusumano et al. (2020)

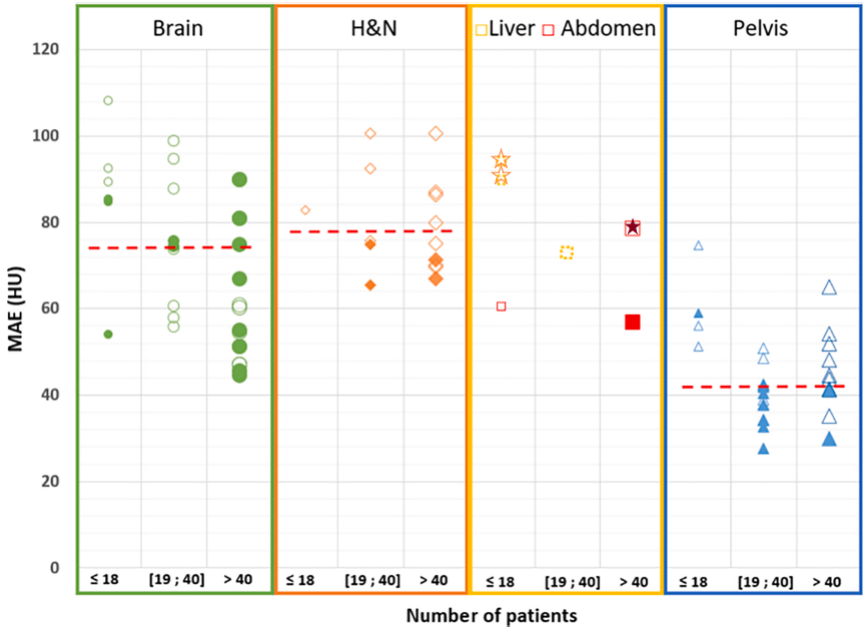


Figure 3.9: Mean absolute error (MAE) results for body structure between reference CT and sCT generated with a deep learning method for studies including the brain, Head and Neck, liver, abdomen, and pelvis. Each marker represent a study result. Full markers represent generator-only models and empty markers generative models with adversarial. Star markers represents the abdominal studies, utilizing low field MR images. Results are divided into three categories: studies including less than 18 patients, studies including 19 to 40 patients and studies including more than 40 patients. Red dotted lines represent the median values. The median values are: 74.2 HU for the brain, 77.9 HU for Head and Neck, and 42.4 HU for the pelvis. Modified from source: Boulanger et al. (2021)

Chapter 4

Experiments: Materials and Methods

4.1 Overview

In this work, we investigated different methods to improve current MR-based sCT generation methods using state-of-the-art generative adversarial networks: pix2pix, CycleGAN, CUT. The following set of experiments was performed:

- Comparing architectures trained in aligned (pix2pix) and non-aligned ways (CycleGAN, CUT).
- Investigating the influence of different preprocessing methods (Nyul, N-peaks).
- Enhancement of results' spatial consistency by adjusting the network configuration (2D, pseudo3D).
- Testing different GAN training objectives (LSGAN, WGAN-GP).
- Estimating the influence of the loss function on the generated results (per-pixel loss versus perceptual loss).

The robustness of the trained networks was analysed using the clinical dataset prepared specifically for this study. This chapter describes the implementation details of the preprocessing pipeline, from the patient selection strategy to the adapted normalisation approaches, as well as the challenges to overcome. The quality of sCT generation was assessed using per-pixel metrics, i.e. mean absolute error (MAE), mean square error (MSE) and peak signal-to-noise ratio (PSNR), as well as perceptual metrics, i.e. structural similarity index (SSIM) and Fréchet-Dirichlet distance (FID). The dosimetric accuracy of the best performing models was estimated comparing the dose distribution of MRgRT treatment plans calculated from sCT and original dCT images using dose-volume histogram parameters to allow assessment of the clinical applicability of the DL methods. The schematic representation of the entire study flow can be seen in Figure 4.1.

4.2 Imaging Data Acquisition and Selection

Imaging data of 76 patients with abdominal tumour treated with MRgRT at USZ was collected retrospectively after approval by the local ethics committee. The ages for these patients ranged from 30 to 85, with a median age of 61; 48 patients were male and 28 were female. The exclusion

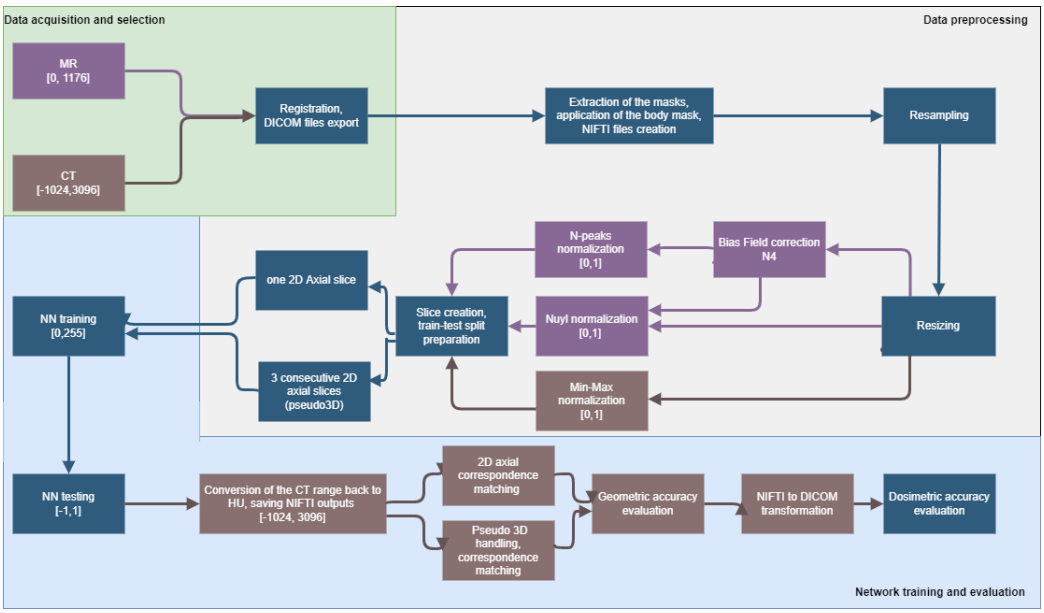


Figure 4.1: Outline of the developed study process. The main stages of the study include: data acquisition and selection; data preprocessing; network training and evaluation. All stages are organised into several steps, which are described in detail in Chapter 4. The steps shown in purple refer to MR images, brown - to CT images, and blue - to both image modalities.

criteria for the study were age below 18 years and the presence of devices in the abdomen that could cause additional artefacts on the images. Patients with kidney stones were included in the study. All data was anonymised. The aim was to obtain imaging data used in the real MRgRT treatment cycle that had not previously been processed or modified.

For each treatment cycle of every patient, a pair of co-registered MR-CT images acquired during the treatment planning was available. This resulted in 93 co-registered volumes as some patients had been treated more than once and had cancer in different organs. Treatment cycles of all patients were performed without the use of a contrast agent. MRI scans were acquired using true fast imaging with steady-state precession (TRUFI) pulse sequence. CT scans were acquired using SIEMENS scanner with 120 KVP. The deformable registration was performed using commercial image registration software, Velocity AI 3.2.1 (Varian Medical Systems, Palo Alto, CA), to align acquired CT images to MR images. The MR pixel spacing was set to 1.5mm and then automatically adapted to the field of view used for each patient with values ranging 1.49 – 1.63 mm. The resolution details of MR and co-registered to it CT images could be found in Table 4.1.

The data was further analysed for the presence of co-registration artefacts. Due to differences in the acquisition process, poor quality of slices at the edge of the field of view, device settings, and moving air pockets, most of the artefacts were present in the first and last slices for each patient. The examples are demonstrated in Figure 4.2 However, the number of corrupted slices varied from treatment to treatment and comprised at least the first and the last 10 slices in the z-direction. To enable an automated procedure for exclusion of the corrupted slices, at first, tumour sizes were examined. Then, the structure sets of delineated tumour(s) and a number of organs for every treatment were exported in DICOM format. In case of multiple tumours or version of PTV delineation, the largest tumour volume was considered (see examples in Figures A.5, A.6). Statistics of the PTV sizes in the axial direction are shown in Figure 4.3 as well. After careful analysis of the data, only 20 slices around the center of tumour were used for the study in order

Pixel Spacing, mm	Slice Thickness, mm	Matrix Size, pix	# Axial Slices	# Treatment Volumes
1.6304 * 1.6304	3	276 * 276	80	**84**
1.4970 * 1.4970	3	300 * 334	144	**3**
1.4957 * 1.4957	3	234 * 234	144	**2**
1.5037 * 1.5037	3	152 * 266	88	**2**
1.4999 * 1.4999	3	310 * 360	88	**1**
1.4970 * 1.4970	3	300 * 334	88	**1**
			Total	**93**

Table 4.1: The resolution of acquired and co-registered MR and CT volumes

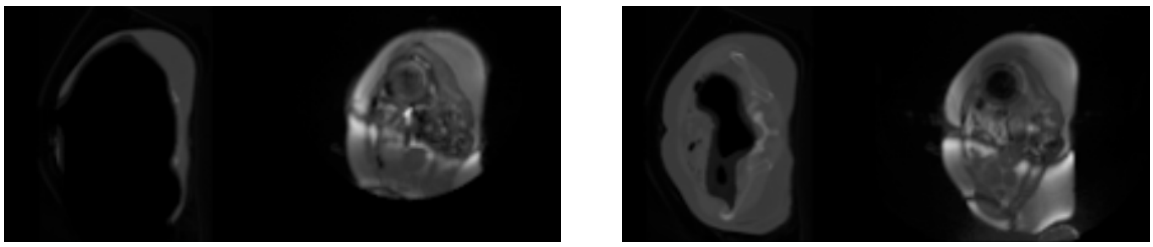


Figure 4.2: CORRUPTED SLICES. Examples of co-registration artefacts at the beginning of the field of view in Z-direction

to maintain high image quality and consistent procedure for selection of the field of view. This strategy made it possible to completely cover the largest tumour and at the same time obtain good quality slices. It resulted in a total number of 3720 slices.

To facilitate the training process and avoid bias in the evaluation of network performance, the entire data set was split into train, test, and validation sets according to the following principles:

- The dataset was split on a treatment-basis into training (80%) and testing (20%) (see Table 4.2, Figure A.13).
- Non-overlapping groups: a patient could only be either in the train or test set (restriction based on the "patient id").
- Stratified folds: each set contained approximately the same percentage of samples for each cancer site as the entire set.
- Restriction of multiple treatments: a patient who had undergone multiple treatments could only be in the train set.
- The validation set consists of 20% of the training set identified with the same conditions (non-overlapping patient groups stratified by cancer area).

The functionality was implemented using the module *sklearn.model_selection* and a few specific manual adjustments.

4.3 Data Preprocessing

Data preprocessing has a significant impact on the quality of the sCT generation tasks, with differences in the individual pixel intensities leading to differences in the generated electron density map and hence differences in the accuracy of the overall irradiation distribution.

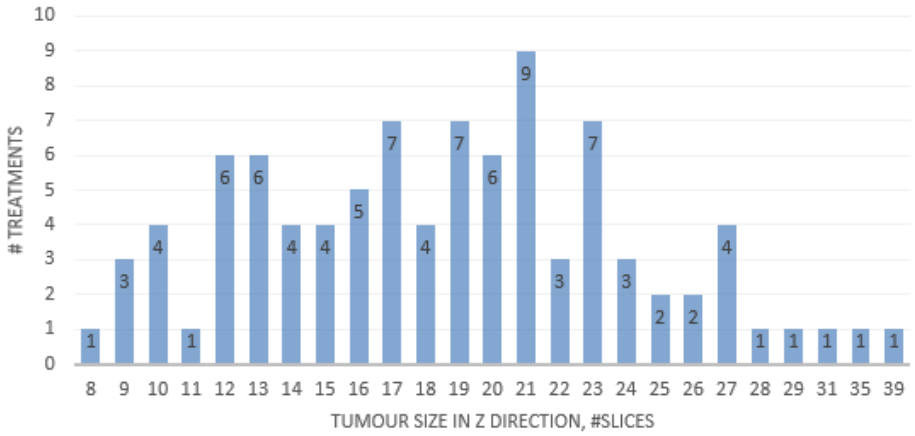


Figure 4.3: Size of the tumour(s) in Z-axis slices. The size of tumour in Z direction in mm could be calculated by multiplying the number of slices by the slice thickness (3 mm)

Cancer area	Train set, treatments (patients)	Test set, treatments (patients)	Total, treatments (patients)
LIV	30 (26)	7 (7)	37 (33)
ABD	16 (12)	3 (3)	19 (15)
ADR	15 (11)	3 (3)	18 (14)
KID	8 (7)	2 (2)	10 (9)
PAN	7 (7)	2 (2)	9 (9)
Total	**76 (59)**	**17 (17)**	**93 (76)**

Table 4.2: Train and test set separation, represented by the cancer areas

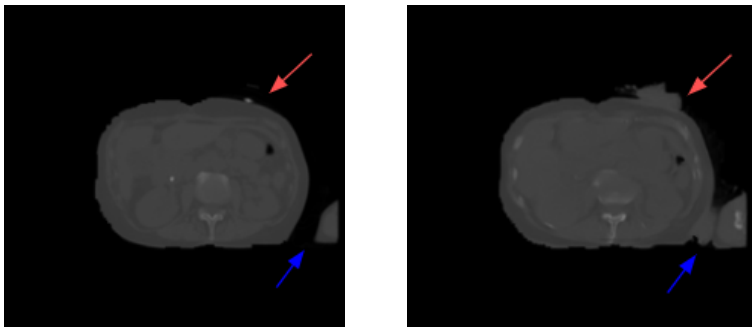
The following steps were performed to remove artefacts in the acquisition, reduce the intensity variations of the same tissues between different volumes, and prepare the inputs for the different network configurations (see Figure 4.1):

- Images conversion from DICOM to NIFTI format. Extraction of body mask and its further application. Extraction of delineated set of organs for tissue-based intensity normalisation. Extraction of the tumour masks for the further FOV adjustment.
- Resampling of the 3D images to match the most common voxel size.
- Resizing of the 3D images to match the common matrix size (cropping or padding).
- Bias field correction for MR images to compensate for the gain variation.
- Nyul intensity normalisation for MR images.
- N-peaks intensity normalisation for MR images. Both intensities are applied to match the ranges in inter-patient settings and to avoid domain shift while NN training.
- Min-max normalisation and intensity clipping for CT images to have a consistent range of values and improve network performance in the bone area.
- Preparation of training, validation and test splits for different experimental settings.

Images conversion from DICOM to NIFTI format. The first essential step was to convert all image files from DICOM to NIFTI format. The advantages of DICOM include interoperability between different software and modality features, performance, and flexibility. The disadvantages include that it requires a significant amount of storage. Whitcher et al. point out that DICOM is not very efficient for image and signal processing (Whitcher et al., 2011), while the NIFTI format is simpler and takes up less space. All the required metadata from the DICOM headers was stored in CSV files. The conversion of the volumes was done using *dicom2nifti* and *nibabel* packages.

At this stage, the binary masks of the delineated tissue sets were also extracted. For this purpose, the DICOM structure file was used, which contains names of all previously delineated tissues defined as Region of Interest (ROI) in the DICOM definitions. The ROI Contour Module is employed to define the ROI as a set of contours. Each ROI contains a sequence of one or more contours, where a contour is either a single point (for a point ROI) or more than one point (for an open or closed polygon) NEMA (2016). After analysing all the delineated ROIs, the decision was reached to use the liver and fat as reference tissues for intensity normalisation and the background, as three reference points are required. The liver was chosen because it was delineated for all patients and it has one of the most homogeneous intensities on MR images among the abdominal organs, with the exception of the urinary bladder. The tag *"liver"* was extracted from the names of *ROIContourSequence*, and its coordinates for each volume were found in the attribute *ContourSequence* for further matching with the original DICOM 3D volume. Fat tissue was selected because it contains a lot of hydrogen, which is the lightest tissue and is present in nearly every patient. In addition, the fat masks were successfully used for tissue-based intensity normalisation and further sCT generation by Farjam et al. (2021) and Hou et al. (2021). Fat mask extraction was based on knowledge of the HU ranges in the CT and additional manual inspection of the population: all pixels with intensities in the range [-160;-60] were assigned to fat. For the more precise selection of the tissue *binary opening* operation were utilised from *scipy.ndimage* package. Further details and limitations of the approach for mask selection are described in the discussion (see Chapter 6). Extraction of the tumour mask with the largest delineated volume or a multiple thereof was performed using the above-mentioned *ROIContourSequence* names containing the tag *"PTV"*. In order to perform the resizing step further, 40 high quality layers around tumour centre was defined on the Z-axis.

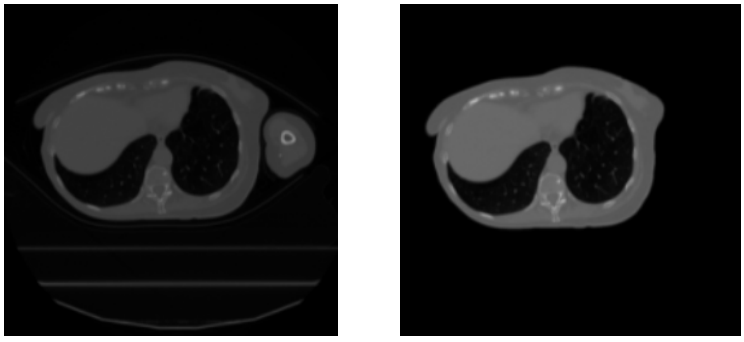
In addition, the binary body masks were further extracted, which were required to focus the histogram-based intensity normalisation and neural networks on the anatomical component of interest. Initially, body masks were extracted from the delineated body contour (tag *"skin"* was used from the names in *ROIContourSequence*). However, the results of the first round of neural network training showed that the manually delineated contour is sharp and contains arms that are more prone to imaging artefacts as they are close to the device boundary (see Figure 6.2a). The results produced by the NN outputs in the experimental round contain significant blurring in the body contour of sCT, as shown in Figure 6.2b. The standard approach to remove unwanted contours using the morphological operations (*binary dilation, binary erosion*) was not powerful enough to remove the arms from the image body contours. Therefore, an approach based on the choice of the largest outlined contour was developed. The proposed solution worked through the functions *findContours* and *drawContours* from *cv2* package. The algorithm involved finding all contours in the manually delineated mask and considering the contour with the largest area, which is the desired body contour not including the arms. This was coupled with the preceding iteration of *binary opening* and three iterations of *binary erosion* to separate the arm contour from the body contour if it is too close to the body, and followed by the reverse operations of *binary dilation* and *binary opening* to smooth the contour. The results obtained showed the smoothing of the contour as well as the exclusion of the arms and the associated artefacts (see Figure 4.7b). Consequently, the proposed algorithm was applied and all variations in the background of both modalities were set to 0 for MR and to -1024 for CT; only the pixels related to the anatomical



(a) Masked body input in trial round of NN training

(b) Output of NN training (CUT)

Figure 4.4: STANDARD MORPHOLOGICAL-BASED BODY CONTOUR MASK APPLIED. The red arrows show the artefacts caused by the marker applied to the body contour as well as by the sharpness of the border; the blue arrows show the artefacts caused by the presence of the hands on the image after training with the old body mask



(a) Unprocessed original CT slice

(b) CT slice after area-based body contour mask applied

Figure 4.5: PROPOSED AREA-BASED BODY CONTOUR MASK. The images show how the proposed method works to exclude the arm (b), which is very close to the body contour in the original image (a)

component remained intact.

The complete set of extracted masks is shown in Figure 4.6. All masks were saved as NIFTI files to allow for further resampling.

Image resampling. Resampling is an integral preprocessing step in medical image analysis. Real-life measurements rely on real-world voxel size. Measurements taken from different images should be comparable(Thévenaz et al., 2000). In order not to influence the learning parameters by variations in voxel size, both MR and dCT slices were resampled to the most commonly represented voxel size in the data: [1.6304 mm,1.6304 mm,3 mm] (see Table 4.1). Resampling of all images was carried out by first-order (linear) spline interpolation using the function *resample_to_output* from *nibabel* package.

Image resizing. When the image was converted to a larger voxel size, the matrix size of an image decreased. To allow the NNs to have the same size of inputs, each axial slice of the image volumes was cropped or padded to the matrix size as following: X=256, Y=256. When cropping, the same number of pixels in one image dimension was trimmed. When padding, the same

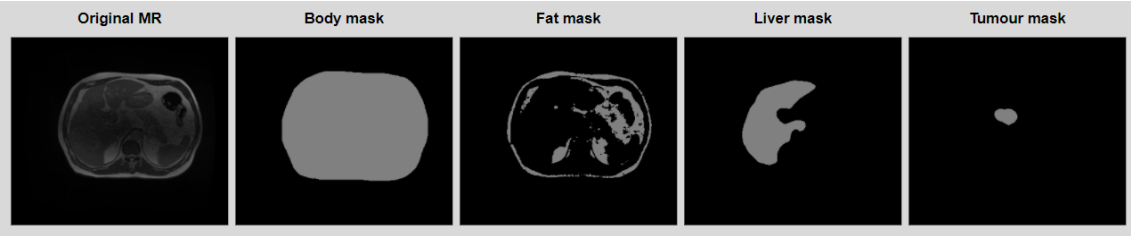
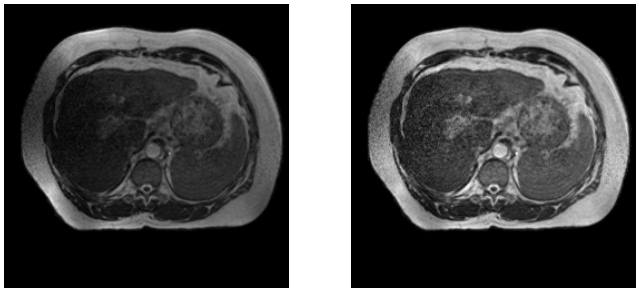


Figure 4.6: Original MR slice and set of masks extracted for it



(a) Before N4 applied (b) After N4 applied

Figure 4.7: BIAS FIELD CORRECTION. The images demonstrate how the N4 algorithm helps to eliminate some of the intensity inhomogeneity. The most obvious effects are observed in the fat and liver tissues

number of pixels was populated with the modality background intensity value. The algorithm is based on the implementation of *crop_or_pad* function from the Ludwig AI [1]. In each treatment cycle, 40 slices were taken in the Z-direction, as mentioned in Section 4.2. This was done by taking the 20 slices around tumour centre. In the event that the ring of 20 slices around the tumour centre touched the first or last 10 slices acquired for a treatment cycle, the centre of the ring was shifted so that it did not touch the first or last 10 slices and still contained 40 slices. In 4 treatments, the shift of the ring centre was performed manually, as artefacts appeared in more than 10 edge slices.

Bias field correction for MR images. MR images are characterised by artefacts during acquisition caused by the inhomogeneity of the magnetic field, the electrical properties of the tissue and the poor uniformity of the coils. To correct for non-linear effects on intensity that vary spatially in an automatic manner, the N4 bias field correction algorithm was applied.

N4 is a variant of the popular retrospective bias correction algorithm N3 (nonparameteric nonuniform normalisation). Based on the assumption that the distortion of the low-frequency bias field can be modelled as the convolution of the intensity histogram by a Gaussian, the basic algorithmic protocol is to iterate between the deconvolution of the intensity histogram by a Gaussian, the reassignment of intensities and the subsequent spatial smoothing of this result by a B-spline modelling of the bias field itself. The changes and improvements over the original N3 algorithm are described in the paper by Tustison et al. (2010)

Refer to Figure 4.7 for examples of MR slices before and after the application of the bias field correction.

Nyul intensity normalisation for MR images. The lack of absolute tissue intensity, especially in the inter-patient setting, is another delicate problem when working with MR images. There are many heterogeneous tissues in the abdomen, which makes it difficult to apply machine learning

[1] https://github.com/ludwig-ai/ludwig

algorithms. In addition, contrast variability is an additional limitation in inter-patient studies. Both of these challenges motivate the use of the Nyul intensity normalisation algorithm (proposed in Nyúl et al. (2000)), one of the few algorithms that can be employed in the abdominal region.

Turning to the details of the normalisation technique, this algorithm tackles the normalisation problem by examining a standard histogram for a set of contrasting images and comparing the intensity of each image to the standard histogram. Many sources refer to this algorithm as a piecewise (affine) histogram-based normalisation, where the standard histogram is learned by demarcating predefined landmarks of interest.

For this study, the Nyul algorithm was adopted from (Reinhold et al., 2019) implementation. The landmarks was defined as intensity percentiles at 2,10,20,. . . ,90,98 percent. The standard scale set to have a predefined range $m^s_{\text{min}} = 0, m^s_{\text{max}} = 100$. The intensity values of the set of MR images I_i were further mapped via following linear map:

$$\tilde{I}_i(\mathbf{x}) = \left(I_i(\mathbf{x}) - m^i_2 + m^s_{\text{min}}\right)\left(\frac{m^s_{\text{max}}}{m^i_{98}}\right) \tag{4.1}$$

which took the intensities of I_i to the range $[m^s_{\text{min}}, m^s_{\text{max}}]$ excluding outliers. Then, deciles for a new image $\tilde{I}_i$ were calculated. This was done for every image $I_i \in I$, where the mean of each calculated value was fixed as the learned reference point for the standard histogram. Thus, for every $n \in \{10, 20, \ldots, 90\}$, the standard scale landmarks were set to:

$$m^s_n = \frac{1}{K}\sum_{i=1}^{K} \tilde{m}^i_n \tag{4.2}$$

After the standard histogram was computed, the set of percentile $\{m_2, m_{10}, m_{20}, \ldots, m_{90}, m_{98}\}$ was learned for every new image. These values were then used to segment the image into ten non-overlapping deciles, which were further identified as $D_{i,j} = \{\mathbf{x} \mid m_i \leq I(\mathbf{x}) < m_j\}$, where $i, j \in \{2, 10, 20, \ldots, 90, 98\}$ and restricting j to equal the next value in the set greater than i. Then, piecewise linear mapping of the intensities, which were associated with these deciles to the corresponding decile on the standard scale landmarks, were performed. The normalised image was then defined as

$$I_{\text{nu}} = \bigcup_{i,j\in\{2,10,20,\ldots,90,98\} i\neq j, i\leq j+10} \left(\frac{I(D_{i,j}) - m_i}{m_j - m_i}\right)\left(m^s_j - m^s_i\right) + m^s_i. \tag{4.3}$$

Finally, the outliers were clipped to the range [0,1]. Refer to Figure 4.8 for examples of MR slices before and after the application of the Nyul normalisation.

N-peaks intensity normalisation for MR images. N-peaks is a normalisation technique recently developed by Wallimann et al. (2022) at USZ. It is based on the idea that two masked tissues and the background should have identical intensities across all images. This enables a consistent physiological interpretation of the normalised images. The N-peaks normalisation is built similarly to the Nyul normalisation method and approaches the normalisation task in two steps: in the first place, it learns a standard histogram consisting of the N homogeneous tissue intensity peaks; in the second place, the intensities of the individual images are mapped onto the standard histogram. Tissue and body masks are required as prerequisites for the method.

Turning to the implementation details in this study, firstly, to find homogeneous areas for each tissue mask provided, a gradient image was calculated for each 3D image volume, showing the degree of local intensity change in each voxel. Then, for each voxel, a neighbourhood of voxels with a distance of a certain radius or less from the voxel was selected. The default radius parameter for 3D images were set to take into account the voxel itself and its 18 neighbours with connectivity 2. Furthermore, in this selection, both the maximum value and the minimum value

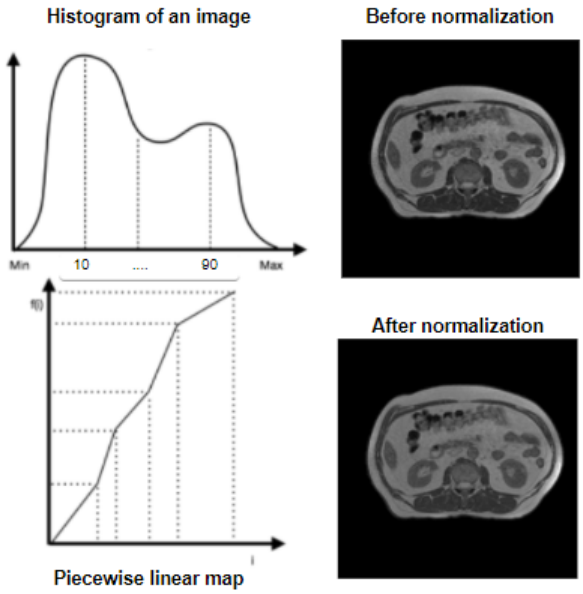


Figure 4.8: Nyul normalisation. The changes in intensities after Nyul normalisation are apparent in the muscle tissue near the spine

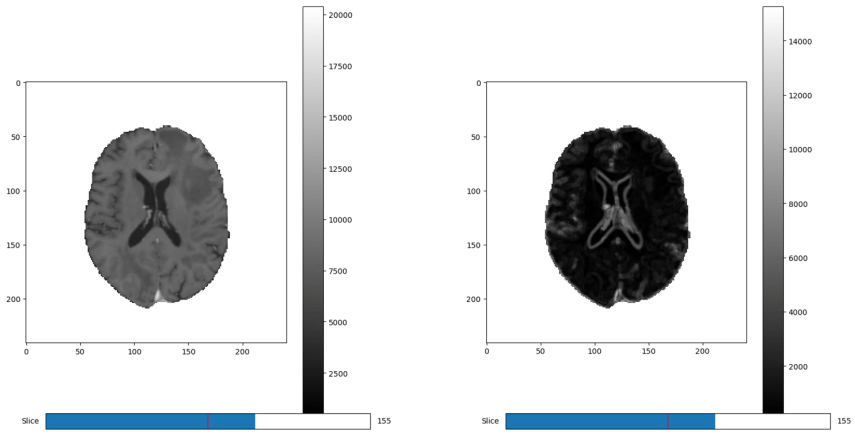
of the intensities were calculated. The gradient was then defined as this local maximum minus the local minimum. The example gradient image is shown in Figure 4.9.

Further on, the algorithm was applied on the level of the given tissue masks (liver, fat) and background: each voxel was stored together with its intensity and its gradient value in the sorted by gradient value flat array. The next steps were related to the main component of the algorithm, which is the interplay between the two values (see Figure 4.10). The voxels, which were closer to the minimum gradient value of a tissue, were then considered as a part of the homogeneous region, while the voxels, which were closer to maximum gradient value, were discarded from the normalisation due to its possibility contains artefacts. The proximity measure was based on the difference between the Jensen-Shannon distance to the most homogeneous and the most heterogeneous snippet (see Figure 4.11). The Jensen-Shannon distance between two probability vectors p and q were defined in the following way:

$$JSD = \sqrt{\frac{D(p \parallel m) + D(q \parallel m))}{2}} \tag{4.4}$$

where m is the pointwise mean of p and q and D is the Kullback-Leibler divergence. The intensity peak for each homogeneous area was then determined based on the probability density function. Concluding the first step, the mean intensity value around the peaks for each tissue mask (liver, fat) in each 3D volume was stored in a data frame as a landmark and considered together with the background value as the standard histogram consisting of the N homogeneous tissue intensity peaks.

In the second step, the mean peak intensity of each tissue was treated as the target intensity. The identified target intensities were then utilised as the basis for a linear normalisation in which each baseline image intensity was transformed using the following equation, similar to the previ-



(a) Input image (b) Its gradient

Figure 4.9: N-PEAKS NORMALISATION. The gradient is calculated for each image to find homogeneous peaks in all provided tissue masks

ously described Nyul normalisation equation:

$$I_{\text{nu}} = \bigcup_{i,j \in \{background, liver, fat\}} \left(\frac{I(D_{i,j}) - m_i}{m_j - m_i} \right) \left(m_j^s - m_i^s \right) + m_i^s \tag{4.5}$$

To make the results comparable to the Nyul normalisation technique, after the linear transformation of the intensities, an outlier handling strategy was implemented based on 2 percentiles from each side. Refer to Figure 4.12 for examples of tissue peak intensities detection. After the outlier handling, the data was scaled to the range [0,1] in similar to the Nyul normalisation fashion.

Min-max normalisation and intensity clipping for CT images CT images are characterised by absolute intensity values as well as a wide range of intensities in bone tissue. Bone tissue intensities can vary in the approximate range [200,3096] HU. The wide range of bone tissue intensities results in ROI intensities concentrated in a very small area, whereas there may only be a few bone voxels with intensities exceeding 1200 HU. This may adversely affect the performance of neural networks.

Based on a meticulous manual examination, the decision was taken to clip the maximum intensities of the CT images to 1200 and then to perform a min-max normalisation, scaling dCT images to the range [0,1] as follows:

$$f(x) = \frac{x - min(x)}{max(x) - min(x)} \tag{4.6}$$

where x and f(x) are the original and standardised intensities, respectively, and min(x) and max(x) are the minimum and maximum image intensity values per patient, respectively.

Preparation of training, validation and test splits for different experimental settings. After applying intensity normalisation methods, different subsets of data were generated to experimentally test different research questions (RQs) to achieve the research objective, which are described in detail in the next section.

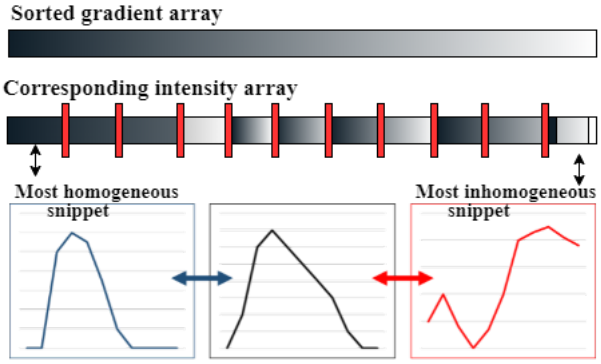


Figure 4.10: The schematic representation of sorted voxel arrays in N-peaks normalisation divided by snippets. The snippet on the left side contains the most homogeneous voxels. The intensities there must therefore correspond to the most homogeneous region in the mask. On the other hand, the snippet at the very right contains the most inhomogeneous voxels. The intensities it contains are therefore not defining a single peak

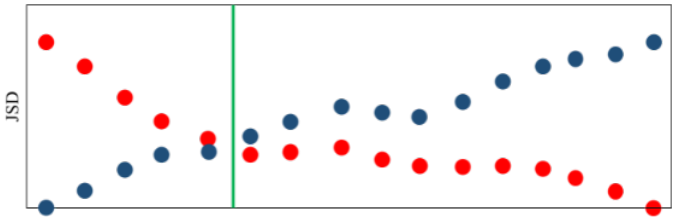


Figure 4.11: JSD in the N-peaks normalisation. For each snippet, the JSD to the most homogeneous snippet is shown in blue. This value naturally starts at 0 on the left, as the snippet has a distance of 0 to itself. This value increases towards the right, as the homogeneous snippet is clearly different from the inhomogeneous snippet. The JSD of each snippet to the most inhomogeneous snippet is simultaneously shown in red. All voxels belonging to the area before the two lines cross (green line) should contain only homogeneous tissue for the given mask

4.4 Experimental Objectives and Key Configurations

In this research, the main objective, which is to improve the current state of the art in DL techniques by studying different architectures and ensembles, was achieved by conducting several experiments with different research questions.

Experiment 1. DL architectures trained on paired vs unpaired data. The first experiment was conducted to analyse the impact of the network architecture and its dataset alignment requirement on the quality of the generated sCT.

RQ 1: Could NN architectures, trained in unpaired fashion, achieve similar performance as architectures, requiring perfectly aligned image pairs in the abdominal area?

To this aim, pix2pix architecture, reproducing the configuration proposed by Isola et al. (see Section 3.2) with U-net generator and PatchGAN discriminator, were trained in aligned fashion, while CycleGAN and CUT, reproducing the configuration proposed by Zhu et al.[2] using a ResNet-based generators and PatchGAN discriminators, were trained in unpaired fashion. The

[2]https://github.com/junyanz/pytorch-CycleGAN-and-pix2pix

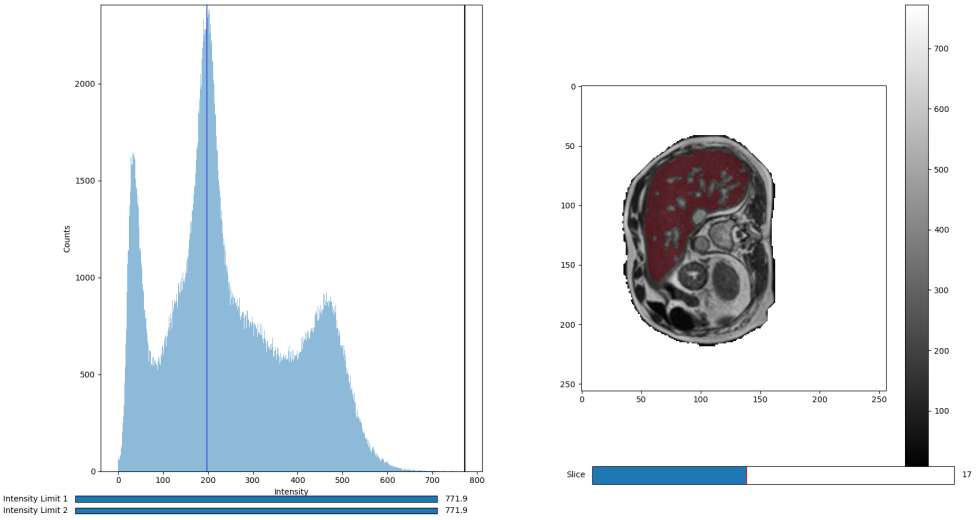
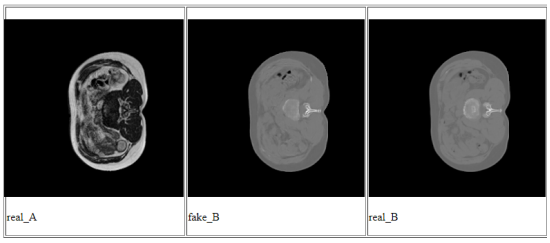
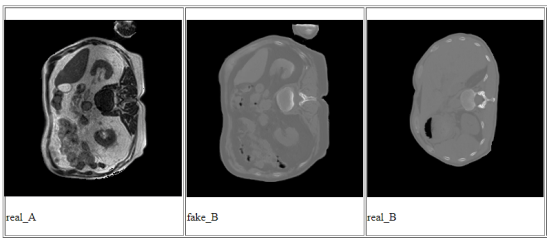


Figure 4.12: N-peaks normalisation. The peak of liver intensity was detected correctly and covers most of the liver area



(a) Paired data training (pix2pix)



(b) Unpaired data training (CUT)

Figure 4.13: EXPERIMENT 1. Examples of the inputs and outputs of the NNs trained in different manners

examples of the different training routines are shown in Figure 4.13. The GANs were trained on the axial 2D slices of the MR and dCT images. The intensities of the MR images were normalised using the Nyul normalisation, and the CT images were normalised using the min-max normalisation, as described in Section 4.3. The detailed configuration sets for each network architecture are available in Attachments A.1,A.2,A.3 and are referenced further as the default configurations.

Experiment 2. Role of the MR image preprocessing. The second experiment examines different approaches to MR data normalisation: histogram based versus biologically motivated intensity normalisation methods.

RQ 2: Could biologically motivated normalisation methods improve the performance of NNs for sCT generation by focusing on specific tissue intensity correction?

To answer this question for each of the NN architectures (pix2pix, CycleGAN, CUT), three different networks were trained, where MR images were initially preprocessed with: Nyul intensity normalisation; Nyul intensity normalisation + N4 bias field correction or N-peaks intensity normalisation + N4 bias field correction. As in the first experiment, NNs were trained on axial two-dimensional slices. The previously mentioned default configurations for each network architecture were exploited.

Experiment 3. Role of the NN input-output channels configuration The third experiment was to quantify the impact of different NN configurations, such as the number of input and output channels.

RQ 3: Could a NN trained with the help of three adjacent 2D slices avoid 3D discontinuities in the area of the abdomen, which is heavily affected by respiratory and peristaltic changes?

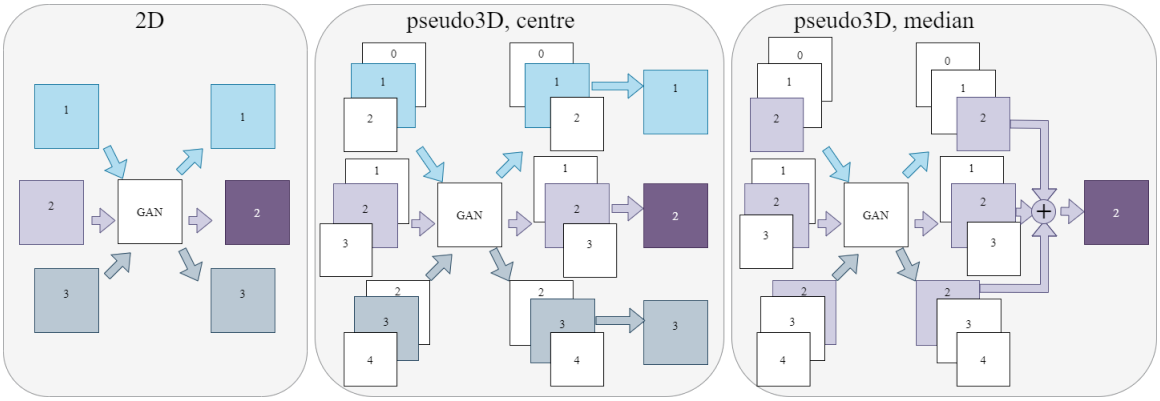


Figure 4.14: Schematic representation of the network configuration studied in Experiment 3, using the end-to-end journey of Slice 2 (dark purple rectangle) of the resulting 3D volume as an example. When testing the network, in the 2D approach, the real MR slice 2 image is passed to the network as a single-channel input, and the sCT image of slice 2 is generated as a single-channel output, which is further directly passed for the creation of a patient's 3D DICOM volume and the evaluation of the quality of the network training. In the pseudo3D approach employed in this study, 3 sequential axial MR slices with stride 1 in the Z dimension are passed as three-channel input, and the sCT images of 3 sequential axial slices are generated as three-channel output. Then, in the Pseudo3D approach, two different strategies are evaluated for combining the results: based on the center slice and the median of the matching sCT slices. In the center approach, for the final evaluation, sCT slice 2 (dark purple rectangle) is taken from the middle slice of the output generated using three MR slices, where slice 2 was in the middle position (the NN pass is indicated by light purple arrows). In the median approach, for the final evaluation, sCT slice 2 (dark purple rectangle) is composed out of all NN passes, where all occurrences of MR slice 2 in the input are taken from the matched output (all light purple output rectangles) and combined using a 1*1*3 median filter with the same weight for each slice position in the output.

To analyse the impact of network configurations on the spatial quality of the generated sCTs for each of the NN architectures (pix2pix, CycleGAN, CUT), two different models were evaluated. The first one was taken for comparison purposes from the first experiment carried out in 2D fashion (single axial slices as inputs and outputs); the second model was trained for this experiment in a pseudo3D fashion (three adjacent axial slices as a NN input and three as an output with one stride in z direction).

Furthermore, two various strategies for combining outputs were additionally evaluated for pseudo3D models. The first considered only the central slice in the 3D output to finally create a 3D DICOM patient volume. In the second case, all output slices corresponding to the same input two-dimensional slices were considered and further merged using a 1*1*3 median filter. The schematic representation of all configurations is explained in greater detail in Figure 4.14. The intensities of the MR images were normalised using the Nyul normalisation, and the CT images were normalised using the min-max normalisation. Default configurations for each network ar-

chitecture were applied, with the only difference being the number of input and output channels.

Experiment 4. Role of the different GAN objectives

The fourth experiment was conducted to determine the GAN objective best suited for sCT generation.

RQ 4: Could different GAN objectives by improving the optimisation process result in a better quality of generated sCTs?

To determine the impact of GAN objectives on output quality, two distinct GAN objectives were evaluated: LSGAN and WSGAN-GP (see Section 3.1). The LSGAN objective used in the pix2pix, CycleGAN and CUT architectures was taken for comparison from the first experiment, as it is the default configuration in the implementation proposed by Zhu et al. [3]. The choice of this objective was motivated by the fact that it forces the generator to sample toward the decision boundary and affects the output performance positively. In order to implement WSGAN-GP objective, which could help to gained improved stability of the optimisation process, especially in terms of the generator optimisation, the default configuration $--gan\_mode$ from the implementation proposed by Zhu et al. [3] was changed to $wgangp$ at first. Then, gradient penalty (GP) computation was employed to three architectures, following the gentle support of the authors of the implementation [4]. All other default configurations remained unchanged, including data normalisation methods (Nyul for MR, min-max for dCT) and the 2D training approach.

Experiment 5. Influence of perceptual loss function

The fifth experiment examines the effect of the perceptual loss function in the pix2pix generator on the quality of SCT in the abdominal region, which is characterised by changes in the shape and position of non-rigid organs and air pockets, as well as barely visible ribs in MR images.

RQ 5: Would using a perceptual loss function in generator instead of a per-pixel loss function help to overcome the known problems in abdomen sCT generation: fuzzy organ boundaries and bone formation errors?

The experiment was inspired by research of Hiasa et al., Lei et al., Kang et al., where the applied perceptual loss functions affected positively generation of synthetic CT images (see Section 3.5). The perceptual loss VGG19 used in this experiment was applied to explore the possibility of achieving visually appealing anatomical detail using the pix2pix architecture. For this purpose, the L1 per-pixel loss (see equation 3.6) was replaced by the VGG19 perceptual loss as follows:

$$\Theta_{G,D} = \arg\min_{G}\max_{D}(L_{GAN}(G_x^y, D_y) + \lambda L_{L_{perceptual}}(G_x^y)) \tag{4.7}$$

where loss weight λ set to 10 and perceptual loss defined as the distance of features extracted by pretrained on ImageNET VGG19 network layers Simonyan and Zisserman (2014) to learn the high-frequency pixel distributions of images:

$$L_{L_{perceptual}}(G_x^y) = \sum_{a=1}^{W}\sum_{b=1}^{H}[\phi_{i,j}(I^y)_{a,b} - \phi_{i,j}(G(I^x))_{a,b}]^2 \tag{4.8}$$

where $\phi_{i,j}$ refers to the feature maps obtained from the j-th Convolution/ReLU pair before the i-th maxpooling layer within the VGG19 network (Takano and Alaghband, 2020).

The components of the SPADE network implementing the VGG19 perceptual loss were used in this experiment (Park et al., 2019). Since the VGG19 network was pretrained on RGB images, the experiment was carried in pseudo3D fashion. All other default configurations remained unchanged, including data normalisation methods (Nyul for MR, min-max for dCT). Afterwards, the synthetic CTs generated using perceptual loss in pix2pix were compared with those generated using per-pixel loss in Experiment 3.

[3] https://github.com/junyanz/pytorch-CycleGAN-and-pix2pix

[4] https://github.com/taesungp/contrastive-unpaired-translation/issues/121

4.5 Evaluation Criteria

As described earlier, 17 MR-dCT image pairs were selected as the test set (see Section 4.2). To evaluate the accuracy of image translation, both geometric accuracy and dosimetric accuracy were assessed.

First, the output values of the generated sCT were linearly scaled to [-1024 HU, 1200 HU], since the intensities of the dCT were initially clipped to this range before being passed to the NN. Then, mean absolute error (MAE), mean square error (MSE), peak signal-to-noise ratio (PSNR), structural similarity (SSIM) between every slice of the sCT generated from the preprocessed MRI and the dCT were calculated as follows:

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|sCT_i - dCT_i \tag{4.9}$$

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(sCT_i - dCT)^2 \tag{4.10}$$

$$PSNR = 20 \cdot \log_{10}\left(\frac{MAX - MIN}{\sqrt{MSE}}\right) \tag{4.11}$$

$$SSIM = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)} \tag{4.12}$$

where i is a pixel within the body contour, N is the total number of pixel within the body contour, and MAX is the maximum pixel value and MIN is the minimum pixel value of the reference image, μ_x and μ_y are the averages of x and y respectively, σ_x^2 and σ_y^2 are the variances of x and y respectively, σ_{xy} is the covariance of x and y. $c_1 = (k_1 L)^2$ and $c_2 = (k_2 L)^2$ are two variables to stabilize the division with weak denominator, L is the dynamic range of the pixel-values, $k_1 = 0.01$ and $k_2 = 0.03$ by default. In addition, the MAE was calculated in two further areas: bone regions and body contour without air pockets. The masks for the bone regions were obtained by thresholding for dCT (x>250 HU). The mask for the body contour without air pockets was obtained by thresholding separately for dCT and sCT (x<-400 HU) and subsequently multiplying the obtained masks.

Additionally, Fréchet inception distance (FID),a measure widely used in studies on GANs proposed by Heusel et al. (2017), was calculated. This metric employs Inception v3 model's last pooling layer to map real and generated images into a feature space (Seitzer, 2020). Then, the Fréchet distance, also referred to a Wasserstein metric, is computed. Intuitively, if the generated images are realistic, they should have similar statistics as real images, and their FID value will be low.

Finally, for the three best-performing models of each architecture (pix2pix, CycleGAN, CUT) the dosimetric accuracy was analyzed based on the MRIdian treatment planning system used for the Co-60 ViewRay system, similar to (Kang et al., 2021). For this purpose, the generated 2D sCT slices were transformed back to 3D DICOM volumes (resized and rescaled back to the original if appropriate). The dose distribution was computed by replacing dCT with sCT images under the same beam parameters as the original dCT treatment plan. The Monte Carlo simulation with magnetic-field correction was performed to calculate the dose distribution. To compare the dose distributions estimated using the sCT and dCT, several dose-volume histogram (DVH) parameters, such as D_2, D_{mean}, D_{95} and D_{98} for PTV and D_2 for one of the OAR (stomach, duodenum, bowel, spinal cord) with the highest prescribed dose as well as D_{mean} for liver were computed (see Section 2.4). In order to evaluate the uniformity and conformality of the sCT-based dose

distribution plan, the absolute difference to original plan in Gy and in $\%$ were calculated in the following way:

$$Difference = \frac{sCT_{dose} - dCT_{dose}}{dCT_{dose}} \tag{4.13}$$

For each DVH parameter the mean and standard deviation (SD) across all evaluated plans was also reported.

Chapter 5

Results

5.1 Experiment 1. DL architectures trained on paired versus unpaired data

RQ 1: Could NN architectures, trained in unpaired fashion, achieve similar performance as architectures, requiring perfectly aligned image pairs in the abdominal area?

In order to evaluate the influence of the different GAN architectures trained on paired and unpaired data, respectively, the results of sCT generation by a 2D pix2pix (trained on paired datasets), CycleGAN and CUT (both trained on unpaired datasets) were compared based on geometric evaluation metrics (see Table 5.1) as well as on visual outcome analysis (see Figure 5.1).

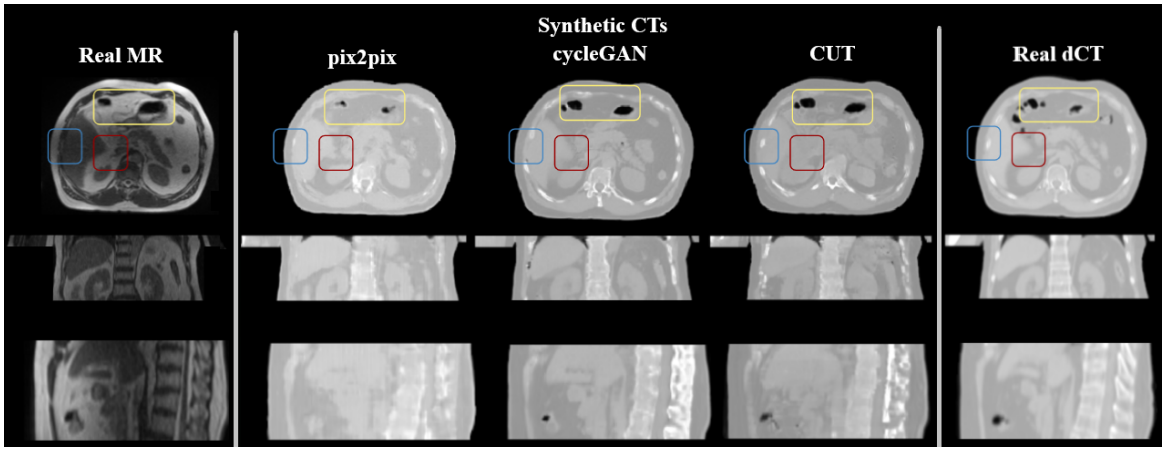


Figure 5.1: Experiment 1. From left to right: original MR image; pix2pix-generated synthetic CT, CycleGAN-generated synthetic CT, CUT-generated synthetic CT (all - fine-tuned); original deformed computed tomography for a patient (high MAE case). From top to bottom: axial, coronal, sagital views. The rectangles highlight some areas of interest for reconstruction quality: yellow - air pockets, blue - ribs, red - liver edge

The synthetic CT images generated through architectures, trained in unpaired fashion, show better quality in dense structures such as the spine and ribs (blue rectangles), with CycleGAN showing the best results in visual outcome. Geometric analysis confirms this finding, showing

	pix2pix		CycleGAN		CUT	
	Baseline	Finetuned	Baseline	Finetuned	Baseline	Finetuned
MAE	75.27 ±21.64	73.78 ±20.91	76.14±19.78	**73.43±20.54**	94.12 ±16.54	90.20 ± 17.85
MAE(excl. air)	54.87 ±12.10	53.67 ±11.84	55.96 ±12.58	**52.27 ±11.73**	65.69 ±11.73	61.99 ±10.53
MAE (bones)	318.40 ±61	320.04 ±57	282.81 ± 40	**269.39 ± 47**	291.01 ±53	302.34 ±54
MSE	2575.45±639	2501.48±618	**2242.89±639**	2280.46±722	3319.79±636	2994.31±598
PSNR	38.26 ±1.04	38.39 ±1.03	**38.90±1.17**	38.86±1.25	37.11 ±0.81	37.57 ±0.85
SSIM	0.982 ±0.010	**0.983 ±0.009**	0.980±0.009	0.982±0.008	0.975 ±0.008	0.976 ±0.008
FID	80.05	78.36	30.66	**29.52**	37.40	42.26

Table 5.1: Results of the first experiments. Baseline models trained with default parameters. In the fine-tuned models learning rate and pool size were changed as following: in pix2pix (*lr*=0.0001, *pool_size=50*), in CycleGAN (*lr*=0.00001, *pool_size=80*) and in CUT (*lr*=0.001, *pool_size=80*)

that CycleGAN has the lowest FID, which is coated to detect contrasts, and MAE in the bone region. However, the ribs produced with CUT are more closely resembling the original MR than those produced with CycleGAN, while ribs on pix2pix sCTs are almost not discernible. It is important to note that bony anatomy has a major impact on the calculation of radiation dose, so accurate identification of bone intensities on CT images can have significant clinical implications.

Additional inconsistencies in the position of air pockets between MR and CT may have a negative impact on the dose estimation. Hence, if any are present due to timing differences in acquisition, these discrepancies are further resolved by manual correction of the electron density maps. As it could be seen from the visual outcome, the location and size of the air pockets (yellow rectangles) are more precise on the synthetic CT images produced by models trained in unpaired fashion (CycleGAN, CUT). Moreover, the quality of the air pocket generation is more stable on synthetic CT generated by CycleGAN and CUT than on the dCT. It demonstrates the superiority of these models over deformable registration methods and DL-based methods, which require perfectly aligned image pairs.

Nevertheless, it can be seen that CUT struggles to produce sharp organ edges (red rectangles), which is also confirmed by the analysis of the geometric indices: MAE and MSE are higher than for CycleGAN and for pix2pix, while PSNR and SSIM are simultaneously lower. Sharp organ boundaries are important for visual image quality and contouring, yet have much less impact on dose calculation than the average pixel intensity difference in the beam path to the tumour. It is evident that the overall "brightness" level between between pix2pix synthetic CT images and the original dCT images is much closer than between CycleGAN and CUT images. This results in low MAE and high SSIM. However, due to the blurring of intensity in the bone area, CycleGAN outperforms pix2pix in the overall geometric estimate.

Overall, answering the research question, both architectures, trained in unpaired fashion, score better in the bone and air pocket areas. Unpaired CycleGAN outperform pix2pix, an architecture trained in a paired manner, on most metrics. However, the average intensity level has room for improvement in architectures trained in unpaired fashion.

5.2 Experiment 2. Role of the MR image preprocessing

RQ 2: Could biologically motivated normalisation methods improve the performance of NNs for sCT generation by focusing on specific tissue intensity correction?

The results of the second experiment show that the different normalisation methods have substantial effects on the performance of models in curtain areas (see Figures 5.2, A.7 for the visual

assessment and Tables 5.2, 5.3, 5.4 for the geometrical metrics). For instance, as would be expected, N peak normalisation applied together with N4 bias field correction gives the clearest liver edge (correspond to red arrow area on dCT), although it gives a slightly worse performance in the spine (blue arrow), compared to Nyul normalisation without bias field correction. Furthermore, one can see that the CUT and CycleGAN models in such a case have drawn an extra rib (green arrow, filled) that has no correspondence on the original MR.

When looking at soft tissue regions, such as the region above the left kidney (yellow rectangle), the sCT generation by both methods, when N4 bias field correction was applied, seems less accurate. The CycleGAN model, trained on the images, which were preprocessed with the histogram-based normalisation method (Nyul) without bias field correction (N4), demonstrates the best results.

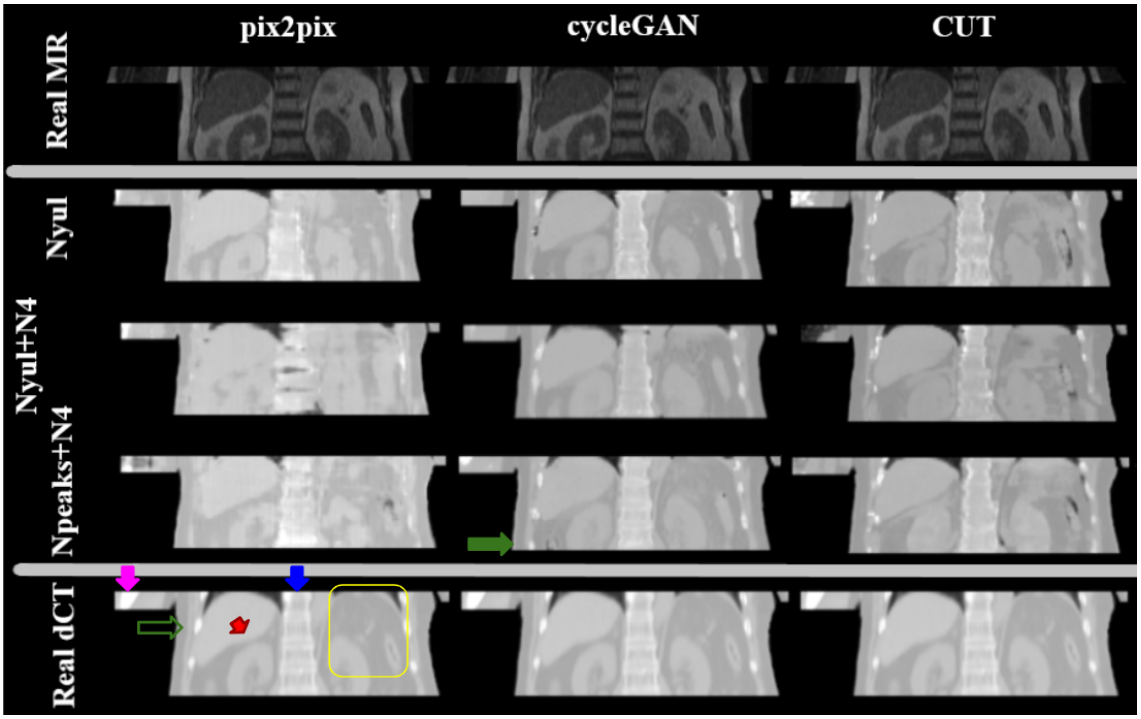


Figure 5.2: Experiment 2. From top to bottom (all - coronal view): real MR image; sCT generated based on: Nyul normalisation applied to input MRI, Nyul and N4 bias field correction, N peaks and N4 original deformed computed tomography for a patient (high MAE case). From left to right: pix2pix (fine-tuned,*lr*=0.0001, *pool_size=50*), CycleGAN (fine-tuned, *lr*=0.00001, *pool_size=80*), CUT (default parameters). The shapes highlight some areas of interest for reconstruction quality in dCT as the example: yellow rectangle - soft tissue above left kidney, blue arrow - spine, red - liver edge, green - rib, purple - arm bone

One of the most valuable results of this experiment is the positive interaction found between CUT architecture and N-peaks preprocessing method. N-peaks, which belongs to biologically motivated normalisation methods, produces gains in almost all metrics for the CUT model, eg. the mean MAE (± standard deviation) decreased from 94.12 ± 16.54 to 84.45 ± 18.96 HU. Visual analysis shows that for complex cases, the N peak + N4 normalisation method used as preprocessing steps for the CUT model helps to solve the average "brightness" problem mentioned in

	pix2pix*		
	Nyul	**Nyul N4**	**N peaks N4**
MAE	73.78 ±20.91	76.33±18.97	**71.72 ±18.04**
MAE (excl. air)	53.67 ±11.84	55.89±10.84	**53.61 ±10.91**
MAE (bones)	320.04 ±57	**307.77 ±56**	321.53 ±53
MSE	**2501.48±618**	2677.38±671	2511.39±632
PSNR	**38.39 ±1.03**	38.10 ±1.04	38.38 ±1.05
SSIM	0.983 ±0.009	0.982±0.008	**0.984 ±0.008**
FID	78.36	**78.32**	83.89

Table 5.2: Results of the second experiments. Pix2pix (*fine-tuned,*lr*=0.0001,*pool_size=50*). Bold metric values show the best performance within an architecture, while blue metric values show the best performance across all architectures

	CycleGAN*		
	Nyul	**Nyul N4**	**N peaks N4**
MAE	**73.43±20.54**	74.43±18.90	77.39±21.33
MAE (excl. air)	52.27 ±11.73	**51.87 ±11.21**	54.46 ±12.65
MAE (bones)	**269.39 ± 47**	274.29±57	273.92±52
MSE	**2280.46±722**	2343.01±660	2378.42±708
PSNR	**38.86±1.25**	38.70±1.15	38.66±1.21
SSIM	**0.982±0.008**	0.981±0.009	0.980±0.009
FID	**29.52**	31.01	32.14

Table 5.3: Results of the second experiments. CycleGAN (*fine-tuned, *lr*=0.00001, *pool_size=80*). Bold metric values show the best performance within an architecture, while bold and blue metric values show the best performance across all architectures

the previous experiment.

Summarising the experimental results, it can be said that normalisation methods have a significant impact on sCT generation outputs with positive attitude in different regions. The best normalisation method have to be chosen for each architecture individually.

5.3 Experiment 3. Role of the NN input-output channels configuration

RQ 3: Could a NN trained with the help of three adjacent 2D slices avoid 3D discontinuities in the area of the abdomen, which is heavily affected by respiratory and peristaltic changes?

Evaluation of the geometrical metrics, provided for the third experiment in Tables 5.5,5.6,5.7, exhibits the superior accuracy of the proposed pseudo3D approach with median slice merging strategy of outputs. The only exceptions are PSNR, which are higher for all models trained in 2D fashion, and FID for CycleGAN and CUT.

Based on visual assessment (see Figure 5.3), the pseudo3D approach with a median slices fusion strategy conveys rough anatomy and preserves structural details in 3D, such as the backbone structure (blue arrow). In CycleGAN, however, the bone structure is smoothed over several lay-

	CUT		
	Nyul	**Nyul N4**	**N peaks N4**
MAE	94.12 ±16.54	91.67± 16.48	**84.45 ±18.96**
MAE (excl. air)	65.69 ±11.73	63.59 ±10.47	**62.07±11.38**
MAE (bones)	**291.01 ±53**	312.23 ±49	297.66 ± 52
MSE	3319.79±636	3047.51±574	**2989.59±635**
PSNR	37.11 ±0.81	37.48 ±0.83	**37.58 ±0.91**
SSIM	0.975 ±0.008	0.975±0.008	**0.979 ±0.008**
FID	37.40	**36.78**	40.49

Table 5.4: Results of the second experiments. CUT trained with default parameters. Bold metric values show the best performance within an architecture, while bold and blue metric values show the best performance across all architectures (here - none)

	pix2pix		
	2D	**pseudo3D, median**	**pseudo3D, central**
MAE	75.3±22	**71.0±20**	72.94 ±20.22
MAE (excl. air)	54.9±12	**51.8 ±11**	53.60 ±10.84
MAE (bones)	318.4±61	293.8±57	**291.50 ±58**
MSE	2575	20323	20913
PSNR	**38.26**	29.63	29.50
SSIM	0.982	**0.984**	0.983
FID	80.05	**70.21**	71.58

Table 5.5: Results of the third experiments. Pix2pix trained with default parameters. Bold metric values show the best performance within an architecture, while bold and blue metric values show the best performance across all architectures

ers, when pseudo3D approach is applied. This is also reflected in the significant increase in the MSE metric, which severely penalises large errors, while MAE assesses the overall impact.

CycleGAN and CUT models are able to capture peristaltic changes better, when trained in pseudo3D fashion, while a certain amount of fuzziness is observed for pix2pix. The CUT model retained the location of the air pocket (green arrow) better than all models.

Answering the research question, pseudo3D models have proven helpful in avoiding 3D discontinuities. The strategy of outputs fusion, based on the median, improves the overall quality of generated sCT images.

5.4 Experiment 4. Role of the different GAN objectives

RQ 4: Could different GAN objectives by improving the optimisation process result in a better quality of generated sCTs?

Figure 5.4 shows resulting sCT slices after inference employing the WGAN-GP objectives in three models. As can be seen, the models trained in unpaired fashion fail to learn CT-specific features, and both CUT and CycleGAN show some air hollows, which is reflected in the metrics shown in Table 5.8. On the positive note, one can notice that CycleGAN and CUT preserve sharp-

	CycleGAN		
	2D	**pseudo3D, median**	**pseudo3D, central**
MAE	76.1±20	**75.5±20**	78.5 ±20
MAE (excl. air)	56.0±13	**55.5±12**	58.4 ±13
MAE (bones)	282.8±40	**275.8±40**	279.5 ±40
MSE	2242.9	22992	24143
PSNR	**38.9**	28.97	28.75
SSIM	0.980	**0.982**	0.980
FID	**30.66**	38.69	37.74

Table 5.6: Results of the third experiments. CycleGAN trained with default parameters. Bold metric values show the best performance within an architecture, while bold and blue metric values show the best performance across all architectures

	CUT		
	2D	**pseudo3D, median**	**pseudo3D, central**
MAE	94.1±17	**84.5±17**	87.8 ±17
MAE (excl. air)	65.7±12	**58.7±12**	61.4 ±12
MAE (bones)	291.0±53	**282.9±52**	288.5 ±50
MSE	3320	27792	29327
PSNR	**37.11**	28.03	27.79
SSIM	0.975	**0.978**	0.977
FID	**37.40**	39.93	39.75

Table 5.7: Results of the third experiments. CUT trained with default parameters. Bold metric values show the best performance within an architecture, while bold and blue metric values show the best performance across all architectures (here - none)

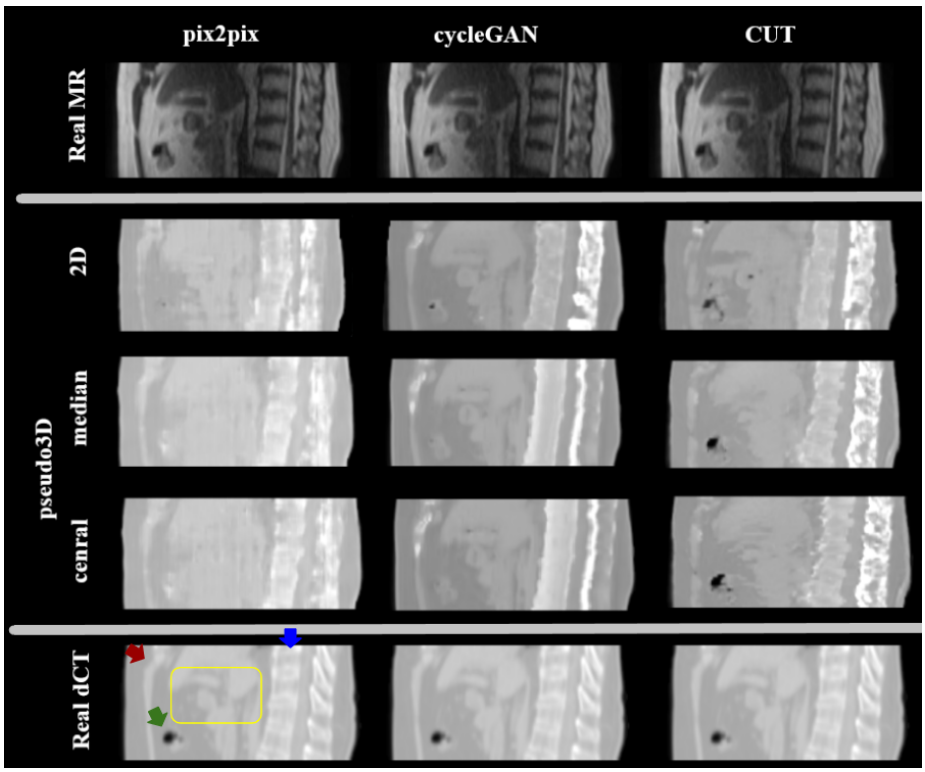


Figure 5.3: Experiment 3. From top to bottom (all - sagital view): real MR image; sCT generated trained in: 2D approach, pseudo3D and merged based on median, pseudo3D and merged based on central slice; original deformed computed tomography for a patient (high MAE case). From left to right: pix2pix, CycleGAN, CUT (all - default parameters, Nyul intensity normalisation). The shapes highlight some areas of interest for reconstruction quality in dCT as the example: yellow rectangle - soft tissue around stomach, blue arrow - spine, red - thoracic wall, green - air pocket

ness of the structural elements, while training with WGAN-GP. At the same time, WGAN-GP-based pix2pix delivers acceptable results, with some additional blurring compared to LSGAN.

Ultimately, the results of the four experiments show that the LSGAN objective is more robust in GAN training and tends to find the optimal solution, seen at least in the current version of the implementation.

5.5 Experiment 5. Influence of perceptual loss function

RQ 5: Would using a perceptual loss function in generator instead of a per-pixel loss function help to overcome the known problems in abdomen sCT generation: fuzzy organ boundaries and bone formation errors?

The sCT produced with the VGG19-based perceptual loss show similar performance to those produced with the L1 loss, both in the visual comparison of the results (see Figure 5.5), where VGG19 contributes with additional blurring in the bone region, and in the comparison of the

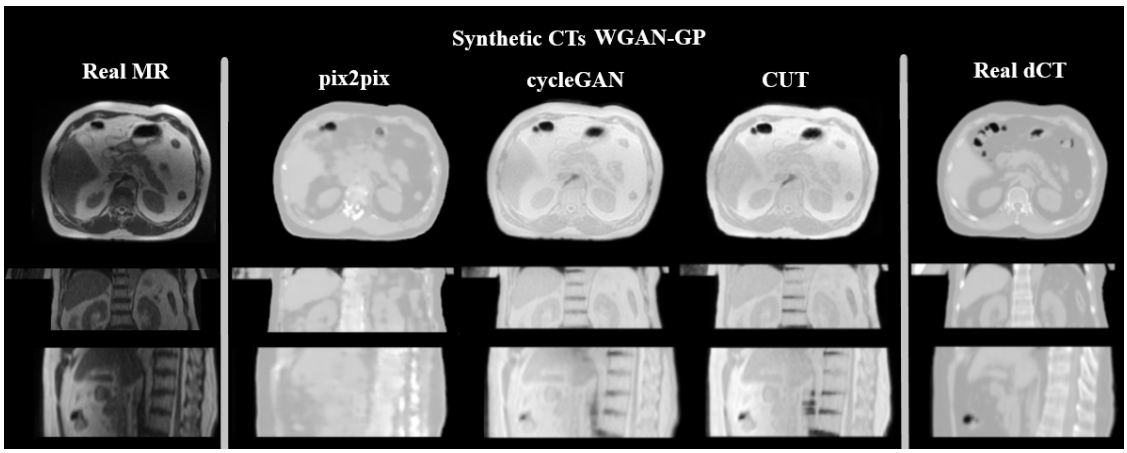


Figure 5.4: Experiment 4. From left to right: original MR image; pix2pix-generated synthetic CT, CycleGAN-generated synthetic CT, CUT-generated synthetic CT (all - default parameters, WGAN-GP); original deformed computed tomography for a patient (high MAE case). From top to bottom: axial, coronal, sagital views

	pix2pix		CycleGAN		CUT	
	LSGAN	**WGAN-GP**	**LSGAN**	**WGAN-GP**	**LSGAN**	**WGAN-GP**
MAE	75.3±22	**71.0±20**	**76.1±20**	149.6±21	**94.1±17**	166.5±24
MAE (excl. air)	54.9±12	**51.8 ±11**	**56.0±13**	132.2 ±15	**65.7±12**	142.4±13
MAE (bones)	**318.4±61**	323.5±55	**282.8±40**	455.1±87	**291.0±53**	448.1±89
MSE	2575	**2388**	**2242.9**	4770	**3320**	5115
PSNR	38.26	**38.61**	**38.9**	35.57	**37.11**	35.27
SSIM	0.982	**0.983**	**0.980**	0.967	**0.975**	0.961
FID	**80.05**	90.54	**30.66**	119.93	**37.40**	92.59

Table 5.8: Results of the fourth experiment. All models trained with default parameters. Bold metric values show the best performance within an architecture

geometric metrics (see Table 5.9). As could be seen, a worse performance in the area of the bones is shown with the default parameters of the training. It can be concluded that the perception loss based on the VGG19, pretrained on ImageNET, does not help to overcome either fuzzy organs or bone formation errors.

5.6 Dosimetric accuracy analysis

As final purpose of the study is to assess the clinical applicability of the DL-based methods, which incorporate the cross-correlation of many parameters in the treatment planning process, the sCT-based dose distribution plans for the best-performing models were further evaluated with the help of MRIdian treatment planning system (see Figures 5.6, A.14, A.15, A.16) for all of the 17 patients from the test set. Model selection was based on geometrical metrics as well as on reconstruction quality in the area of interest: organ boundaries, bones and overall soft tissue intensities. The best performing models include: pix2pix trained in Pseudo3D with LSGAN objective and L1 loss, MR preprocessed with Nyul (see Table 5.5, MAE 71.0±20 HU); CycleGAN trained in 2D with

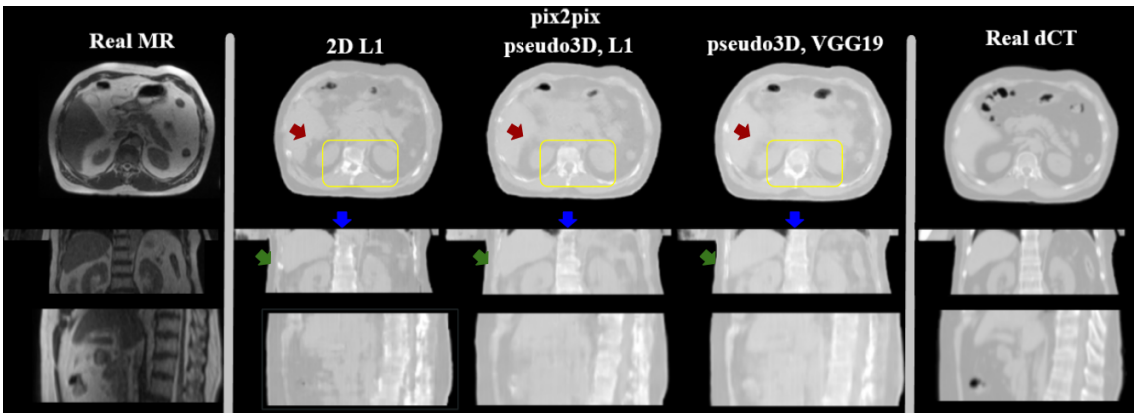


Figure 5.5: Experiment 5. From left to right: original MR image; pix2pix-generated synthetic CT: in 2D fashion with L1 loss, in pseudo3D fashion with L1 per-pixel loss, in pseudo3D fashion with VGG19 perceptual loss; original deformed computed tomography for a patient (high MAE case). From top to bottom: axial, coronal, sagital views. The shapes highlight some areas of interest for reconstruction quality in dCT as the example: yellow rectangle - spine on axial view, blue arrow - spine on coronal view, red - liver edge, green - rib

	pix2pix, pseudo3D	
	L1	**VGG19**
MAE	**71.0±20**	72.5±20
MAE (excl. air)	**51.8 ±11**	53.9 ±12
MAE (bones)	**293.8±57**	330.8±57
MSE	20323	20164
PSNR	**29.63**	29.61
SSIM	0.984	0.984
FID	70.21	**68.91**

Table 5.9: Results of the fifth experiment. Both models trained with default parameters. Bold metric values show the best performance within an architecture

LSGAN objective and L1 loss, MR preprocessed with Nyul (see Table 5.1, MAE 73.4±21); and CUT trained in 2D with LSGAN objective and L1 loss, MR preprocessed with N-peaks and N4 (see Table 5.2, MAE 84.5 ±19). Box-plot analysis showing the dosimetric differences in % and Gy for every DVH indicator (minimum, maximum, median, first and third quartiles) of the three best models is shared between Figures 5.7 and 5.8, respectively. Mean and standard deviation (SD) of differences are reported in the Table 5.10. The representative plans with low and high DVH indicator differences could be found in Figures A.8, A.9.

The dosimetric accuracy analysis shows that, with the exception of a few outliers, the discrepancies between the dose differences of sCT-based and original dCT-based plans are less than 1% for most of the patients, demonstrating that **the proposed DL-based sCT generation methods may be considered clinically applicable for treatment planning in the abdominal area**. Moreover, the model with the smallest differences, CycleGAN, demonstrates high uniformity and concordance of dose distribution plans, with differences for all PTV DVH indicators of less than 0.5% (less than 0.3 Gy) and the smallest standard deviation.

DVH indicator	pix2pix Diff, %	pix2pix Diff, Gy	cycleGAN Diff, %	cycleGAN Diff, Gy	CUT Diff, %	CUT Diff, Gy
PTV Dmean	0.33 (0.27)	0.15 (0.13)	**0.32 (0.21)**	**0.13 (0.10)**	0.40 (0.37)	0.17 (0.17)
PTV D2	**0.33 (0.27)**	**0.18 (0.16)**	0.44 (0.37)	0.20 (0.16)	0.51 (0.37)	0.25 (0.21)
PTV D95	0.53 (0.63)	0.17 (0.19)	**0.38 (0.28)**	**0.13 (0.09)**	0.44 (0.36)	0.15 (0.14)
PTV D98	0.57 (0.61)	0.18 (0.18)	**0.48 (0.41)**	**0.14 (0.11)**	0.53 (0.38)	0.16 (0.11)
OAR D2*	0.57 (0.60)	0.10 (0.10)	**0.56 (0.46)**	**0.09 (0.07)**	0.65 (0.60)	0.10 (0.09)
Liver Dmean*	0.29 (0.30)	0.04 (0.04)	**0.31 (0.25)**	**0.04 (0.04)**	0.30 (0.30)	0.04 (0.04)

Table 5.10: Dosimetric accuracy evaluation of the best performing models. Mean values of dose difference calculated between sCT and dCT for all the DVH indicators considered, calculated based on the absolute dose value differences, which were reported in Gy and %. For each DVH parameter the standard deviation (SD) is reported. Bold values shows the lowest difference across all architectures. * Among the OARs for this study, duodenum, stomach, bowel, spinal cord are considered, only the organ with the highest D2 dose were considered for each patient. While computing the mean and standard deviation for liver Dmean, doses below below 5Gy were excluded

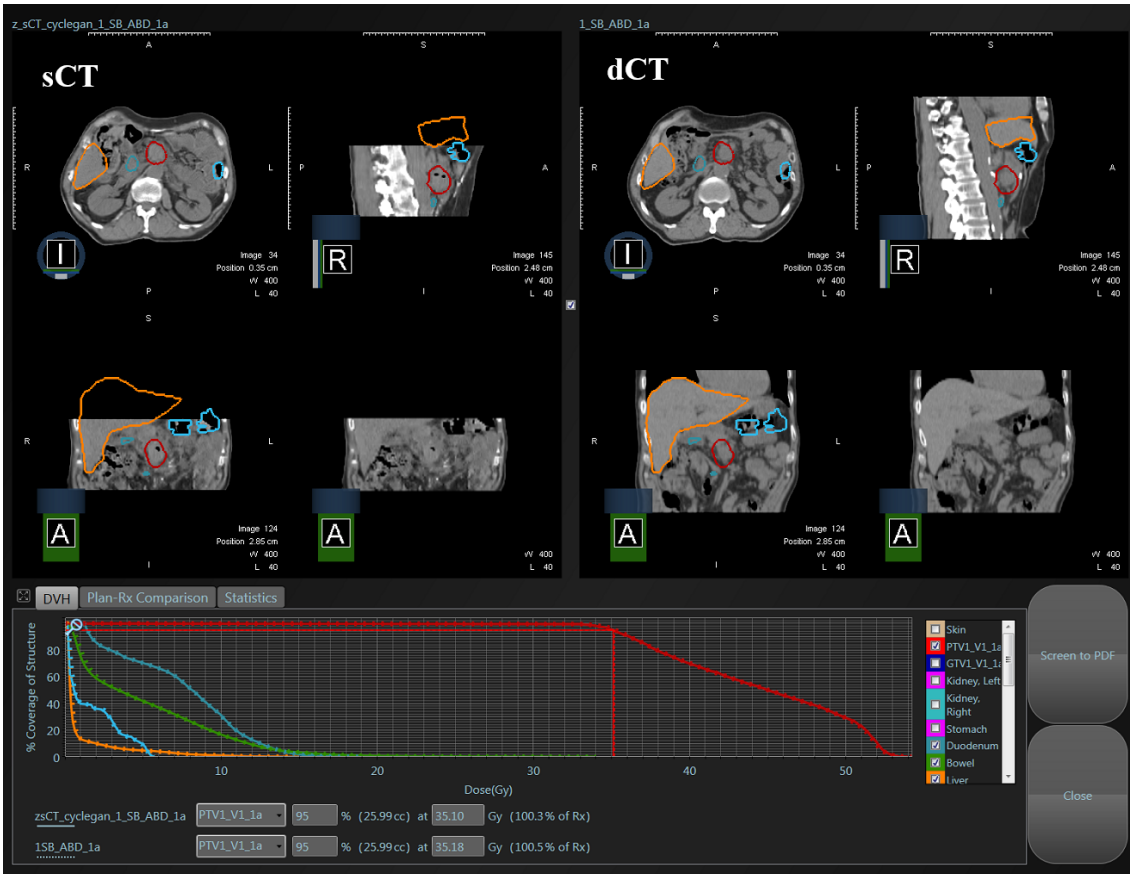


Figure 5.6: Comparison of dose volume histogram for PTV and OARs between sCT generated by CycleGAN (low MAE case) shown on left and original dCT shown on right. Dashed line indicates the result of dCT and the solid line shows the result of sCT methods

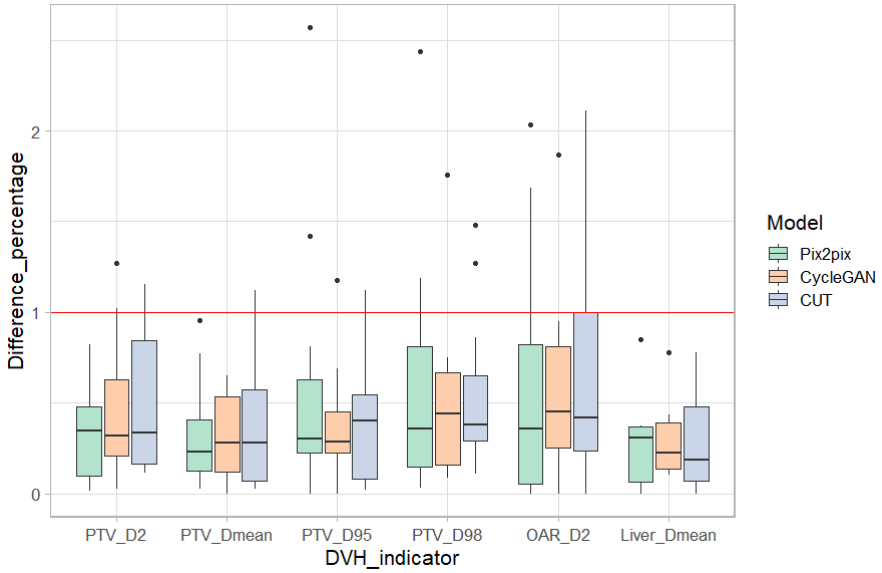


Figure 5.7: The box plot analysis of the DVH differences in %. The red line shows a threshold for clinical applicability. Among the OARs for this study, duodenum, stomach, bowel, spinal cord are considered, only the organ with the highest D2 dose were considered for each patient. While computing the mean and standard deviation for liver Dmean, doses below below 5Gy were excluded

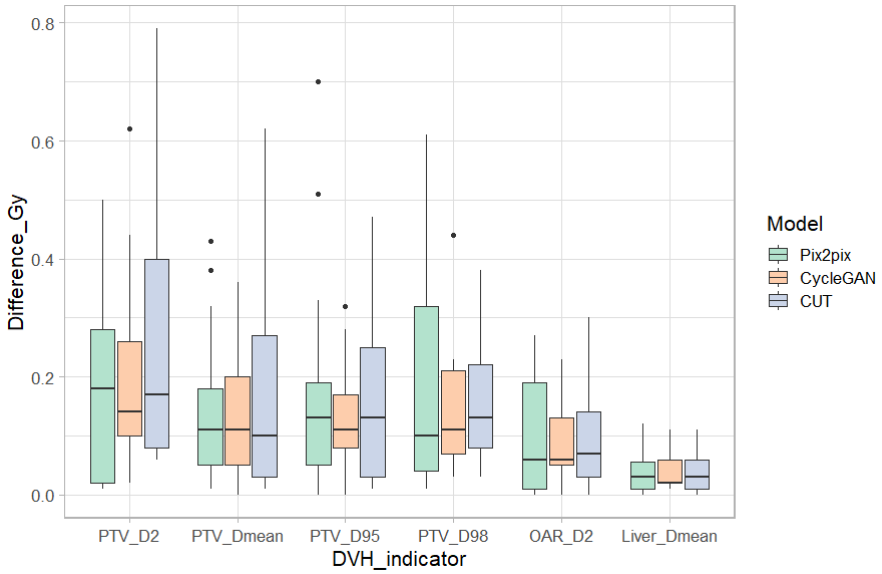


Figure 5.8: The box plot analysis of the DVH differences in Gy. Among the OARs for this study, duodenum, stomach, bowel, spinal cord are considered, only the organ with the highest D2 dose were considered for each patient. While computing the mean and standard deviation for liver Dmean, doses below below 5Gy were excluded

Chapter 6

Discussion

The present study is aimed to devise deep learning approaches that are able to generate realistic sCT images from MR images. The study provides additional insights into their generalisation capabilities and the quality evaluation of sCT-based dose distribution plans, a critical prior for the clinical implementation of new technological advances.

Experimental results show that DL models trained in the unpaired fashion (CycleGAN, CUT) could achieve comparable results to the paired trained pix2pix model, with better performance in generating air pockets and bones and slightly more moderate in generating average intensities of soft tissues. Although it is impossible to state which normalisation method - histogram-based or biologically motivated - is ultimately best for preprocessing across all GAN models, it has been found that the novel N-peaks normalization helps to achieve significantly better quality of sCT generation for CUT models. Moreover, a pseudo3D configuration for neural networks coupled with the median fusion strategy is confirmed as another key component to help avoid three-dimensional discontinuities in an abdominal region highly susceptible to respiratory and peristaltic changes, and LSGAN with L1 loss may be favoured as more robust objective for the GAN training .

In terms of image accuracy, the geometric evaluation of the model performance obtained in this study and the comparison as a reference for the current state-of-the-art (Cusumano et al. (2020), Kang et al. (2021)) in abdominal sCT generated from 0.35 MRI for MRgRT are presented in Table 6.1. As can be seen from the table, both pix2pix and CycleGAN configured in this study outperform the state of the art in all metrics except MAE (bone), which was reported by Cusumano et al. For pix2pix, the mean MAE within the body contour was dropped by 7 HU, which corresponds to an improvement of around 9.8%, compare to the results of our study, while comparing the performance of pix2pix from Cusumano et al. with our CycleGAN architecture corresponds to a mean decrease in MAE by around 5 HU and 6.7%. The mean MAE reported by Kang et al. for the CycleGAN architecture is difficult to compare with the results of this study because MAE, MSE (and those PSNR) were both scaled to the number of voxels within the image, whereas in this study the errors were scaled to the number of voxels within the body contour, which naturally makes the MAE and PSNR numbers higher overall. However, even in such an setting, the CycleGAN architecture used in this study could be seen to significantly outperform the results of Kang et al. with an increase in SSIM of almost 7.6%, which corresponds to a significant image quality improvement. The relative position of this study compared to others using DL-based methods to generate sCT is shown in Figure 6.1. As it could be seen this study outperforms state-of-the art, trained on low field magnetic resonance images. The studies showing better results have been trained on higher field MR images.

A further comparison with the state-of-the-art, based on the dosimetric accuracy, (see table 6.2) reveals that **the best performing model proposed in this study (CycleGAN) outperforms the state-of-the-arts for all DVH indicators**. Furthermore, the lower variability in the estimation

	Our results (73 patients in dataset)			Cusumano et al. 60 patients	Kang et al. 90 patients*
	pix2pix	**CycleGAN**	**CUT**	**pix2pix**	**CycleGAN**
MAE	**71.0±20**	73.43±20.54	84.45 ±18.96	78.71 ± 18.46	58.8 ± 4.4 ***
MAE (excl. air)	**51.8 ±11**	52.27 ±11.73	62.07±11.38		
MAE (bones)	293.8±57	269.39 ± 47	297.66 ± 52	**152.71±30.14**	
MSE	20323?	2280.46±722	2989.59±635		
PSNR	29.63	**38.86±1.25**	37.58 ±0.91		26.3 ± 0.7***
SSIM	**0.984**	0.982±0.008	0.979 ±0.008		0.91 ± 0.01
FID	70.21	**29.52**	40.49		

Table 6.1: Comparison of the state-of-the-art with the results of the our study based on the geometrical evaluation. * NN trained on patients who had pelvic (n = 24), thoracic (n = 24) and abdominal (n = 24) cancer for the purpose of NN generalisability, results provided for abdominal sCT test (n = 6). *** MAE and MSE within the body were scaled to the total number of image voxels in the study by Kang et al., which differs from the method of our study where the error is scaled to a much smaller number of voxels within the body contour

of DVH parameters for sCT plans is observed for all methods proposed in this study, compared to state-of-the-arts.

Moreover, the compatibility of sCT-based plans, in which dose evaluation was performed without correction for air pocket position change, compared to dCT-based plans, in which such correction was applied in advance, shows that DL-based methods could eliminate costly work for the clinicians while still providing dosimetric evaluation of high quality. The example of the DL-produced sCTs, which capture the position of the air pockets well, can be seen in Figures 6.2, A.10.

Although the CUT architecture, which was first applied to generate sCT in this study, shows moderate performance based on the geometrical evaluation with MAE 84.5±19 HU, compared to best performing pix2pix with MAE 71.0±20 HU and cycleGAN with MAE 73.43±20.54 HU, the results of the dosimetric accuracy reveals it high potential for further upgrade. The CUT model has outperformed the results of the state-of-the-art CycleGAN model, proposed by Kang et al. (2021) in all of the comparable metrics (PTV mean difference equals to 0.2 Gy versus 0.17 Gy , PTV D2 is 0.4 Gy versus 0.25 Gy and PTV D98 equals to 0.5 Gy versus 0.16 for the CycleGAN proposed by Kang et al. (2021) and CUT from this research, respectively). Moreover, the additional motivation for further CUT investigation is the fact that the current research showed that the model could be successfully trained in unpaired manner, for which purpose more CT and MR data are available, which are, however, not co-registered, and those, could not be used for pix2pix.

During the implementation different scenarios that might hinder the validity of the study results were observed. Most of them, however, may be addressed and investigated in follow-up studies.

One of the main concerns relates to the dataset formation. In contrast to most of the studies (Cusumano et al., 2020; Kang et al., 2021; Klages et al., 2020) no manual exclusion of 3D volumes was performed due to the low signal or lack of spatial integrity between MR and co-registered CT, as well as significant peristaltic changes, large tumours or bone co-registration artefacts. Moreover, the lower quality images, such as those with a significant bias field (see real MR image in Figure 5.3) or some acquisition artefacts (see Figures A.11, A.12), were retained in the dataset. All of those facts contribute to removing the potential bias of a specific input quality may have on training and generation of a robust model. However, all the CT and MR images used in the study were acquired on the same devices and MR was acquired only with 0.35T magnetic field, which

DVH indicator	**Our results (73 pat. in dataset): Mean difference (SD) : for PTV below 0.5% (below 0.3 Gy)**						**Related work (MR 0.35T)**	
	pix2pix		**CycleGAN**		**CUT**		**Pix2pix Cusumano et al. 60 pat.**	**CycleGAN Kang et al. 90 pat.****
	Diff, %	Diff, Gy	Diff, %	Diff, Gy	Diff, %	Diff, Gy	Diff	Diff, Gy
PTV Dmean	0.33 (0.27)	0.15 (0.13)	**0.32 (0.21)**	**0.13 (0.10)**	0.40 (0.37)	0.17 (0.17)	-0.08 (0.22)Gy*	0.2
PTV D2	**0.33 (0.27)**	**0.18 (0.16)**	0.44 (0.37)	0.20 (0.16)	0.51 (0.37)	0.25 (0.21)	-0.13 (0.3) Gy	0.4
PTV D95	0.53 (0.63)	0.17 (0.19)	**0.38 (0.28)**	**0.13 (0.09)**	0.44 (0.36)	0.15 (0.14)	-0.28 (1.06) %	
PTV D98	0.57 (0.61)	0.18 (0.18)	**0.48 (0.41)**	**0.14 (0.11)**	0.53 (0.38)	0.16 (0.11)	-0.05 (0.23) Gy	0.5
OAR D2***	0.57 (0.60)	0.10 (0.10)	**0.56 (0.46)**	**0.09 (0.07)**	0.65 (0.60)	0.10 (0.09)	-0.04 (0.23) Gy	
Liver Dmean***	0.29 (0.30)	0.04 (0.04)	**0.31 (0.25)**	**0.04 (0.04)**	0.30 (0.30)	0.04 (0.04)		

Table 6.2: Comparison of the state-of-the-art with the results of the our study based on the dosimetric accuracy evaluation. Mean values of dose difference calculated between sCT and dCT for all the DVH indicators considered, calculated based on the absolute dose value differences, which were reported in Gy for all the parameters as well as in percents. For each DVH parameter the standard deviation (SD) is reported. The DVH difference for state-of-the-art is reported in units they were reported in papers.*D50 is reported in original paper. ** NN trained on patients who had pelvic (n = 24), thoracic (n = 24) and abdominal (n = 24) cancer for the purpose of NN generalisability, results provided for abdominal sCT test (n = 6). *** Among the OARs for this study, duodenum, stomach, bowel are considered, only the organ with the highest D2 dose were considered for each patient. While computing the mean and standard deviation for liver Dmean, doses below below 5Gy were excluded. In study by Cusumano et al. Duodenum/Bowel were considered as OARs

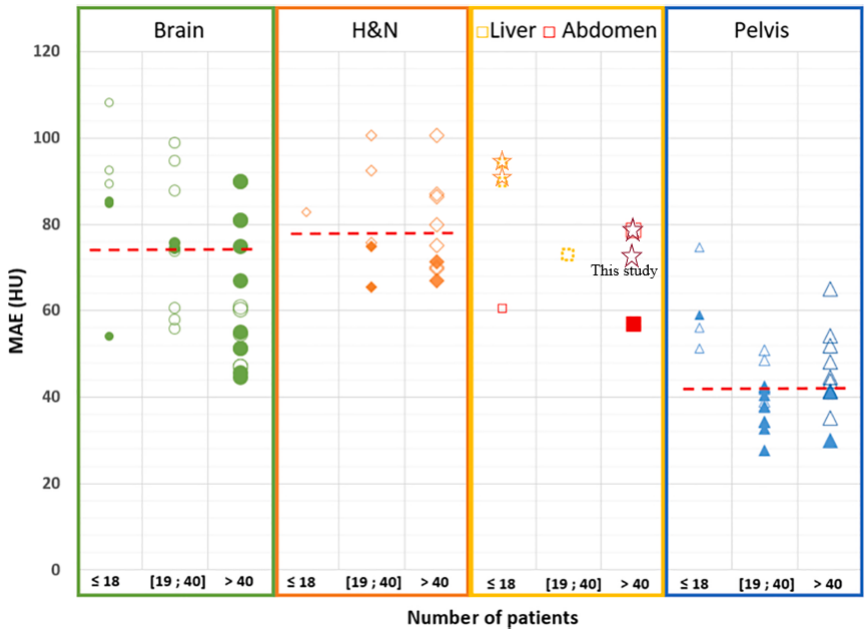
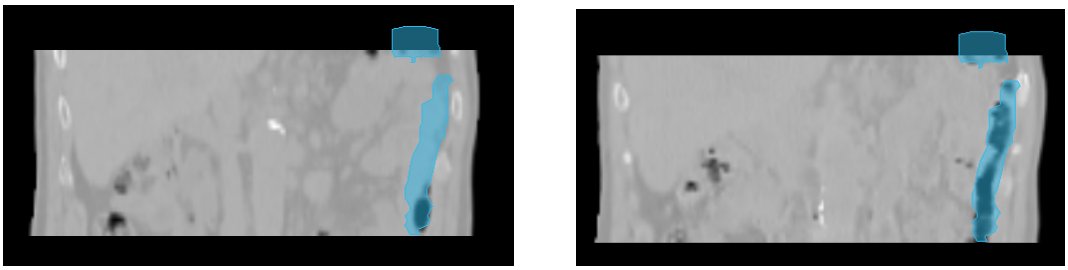


Figure 6.1: Mean absolute error (MAE) results for body structure between reference CT and sCT generated with a deep learning method for studies including the brain, HN, liver, abdomen, and pelvis. Each marker represent a study result. Star markers represent abdominal studies that have been trained under conditions similar to this study (0.35T MR images). The study by Kang et al. was not mapped due to lack of MAE figures within the body contour. It can be seen that there are few studies in the abdominal region, with this study being superior to the state-of-the-art, based on geometrical accuracy evaluation. Modified version taken from: Boulanger et al. (2021)

could cause lower model generalisation capability in case of employment of a more heterogeneous dataset. In order to solve the problem of sample variability, additional data augmentation could be carried out. In addition, patients with metal implants were excluded. Therefore, further investigation of the errors in the DL-based sCT generation specific to the mentioned patient cohort is required.

Alongside the quality of the images produced, an important factor in sCT implementation in clinical practice is the speed of sCT generation and the dosimetric calculation based on it: this is particularly important in procedures such as MRgRT, where the patient is in the treatment position waiting for the plan adjustment to be completed. Due to the time constraints of the research, speed is beyond the scope of this thesis.

Another limitation is associated with the implementation of the preprocessing routine for every model. The study has revealed the importance of MR preprocessing methods in the abdominal region, with a potential sCT generation quality gain within the same DL model of about 10% in MAE mean (see results of N-peak N4 preprocessing (biologically motivated) compared to Nyul preprocessing (histogram-based) method for CUT model in Table 5.4). This result goes in line with the results of the largest studies related to sCT generation in the brain area on about 400 patients by Andres et al. (2020), where the white-stripe normalisation technique, which belongs to the group of biologically motivated normalisation techniques, outperformed the histogram-based normalisation with mean MAE head equals to 78 HU and 92 HU correspondingly. In the human



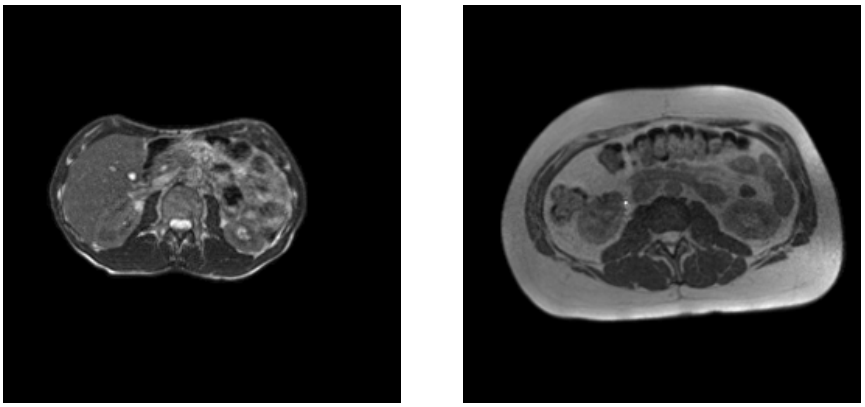
(a) Original dCT (b) Generated sCT

Figure 6.2: AIR_OR STRUCTURE. The example of the original dCT and sCT generated by CycleGAN coronal slices, overlaid with the air pocket delineated on original MR drawn in blue. The position of the air pocket is better captured by the DL-based method for sCT generation

brain, the choice of tissues for normalisation (white matter, grey matter) and their automatic mask delineation based on brain anatomy does not cause additional burden on the validity. At the same time, no studies were found that focused on investigating the best preprocessing routine for generating abdominal sCT. For this reason, the choice of fat and liver tissue as landmarks for N-peaks normalisation in this study could be crucial for evaluating the performance of the method due to the different anatomy of the patients. Some of the patients in the study had almost no fat tissue, only muscle, while in many others the presence of muscle tissue was minimal (examples are shown in Figure 6.3). Therefore, when applied to the different populations, the algorithm may not find the biological landmark for normalisation in some patients and the results may be different. Besides, the fat masks for N-peaks normalisation were calculated based on the intensities of dCT images, which means that the current technique can not be readily applied in practice if CT imaging is completely excluded from the radiotherapy routine. The solution to this hurdle could lie in adapting the various methods for MR-based automatic soft tissue segmentation in the abdominal region, as it has been done with the fuzzy C-means (FCM) clustering algorithm (Hou et al., 2021) or with graph cuts and image derived energies Christensen et al. (2017) or via DL-based approaches (Estrada et al., 2020). Finally, N4-Algorithm was applied as a step in the preprocessing routine, as it was found that N-peak normalisation without it during the pilot trials is struggling to identify correct tissue peaks due to the large variations in intensities of neighbouring pixels within the intensities of a tissue. Therefore, the exploration and fine-tuning of techniques to correcting bias fields is of interest, as they can strongly influence the results of biologically motivated intensity normalisation.

In addition, the elimination or intensification of the CT intensity range clipping in the bone region could be important if the preprocessing workflow for sCT generation would be applied to another clinical task, such as CT reconstruction for orthopedic purposes (Hiasa et al., 2018).

The results of the third experiment, where the pseudo3D training approach showed better performance than the median output merging strategy, especially in terms of the mean MAE value of pix2pix and CUT for whole-body contours, which increased by 6 and 10%, respectively, could be explained by the increase in the network perceptive field for each output pixel. Furthermore, despite its blurring effect, the median filter seems to provide greater benefit by accumulating the information from the largest number of slices compared to the poorer merging strategy based only on the central slice. Since the Nyul normalisation method was used for all architectures for this experiment, it is intriguing to find out if the synergies between the biologically motivated normalisation methods and the network configurations could yield an even higher overall quality scores. The choice of whole slices as the spatial configuration for the experiment was motivated by the expectation that for the network trained in an unpaired manner, especially for CUT, manipulat-



(a) Almost no fat tissue (b) Minimal presence of muscle tissue

Figure 6.3: N-PEAKS NORMALIZATION CHALLENGES. The different physiological conditions of patients in the abdominal cavity make it difficult to select certain tissues for normalisation

ing mainly within the input-output pairs of the generator, would be enhanced by capturing the contextual information from the whole image. Additionally, retaining the full image resolution could decrease the time in the end-to-end flow for sCT-based dose evaluation. The results of the experiment, though, showed the relevance of increasing the size of the receptive field for sCT generation tasks, which is consistent with the results of Klages et al. (2020), where the combination of overlapping patches and multiple views was found to boost performance in the head and neck area. Therefore, researching further configuration into NN operating the abdominal area, which is known for its strong physiological changes compared to the other areas of cancer treatment, is another matter of interest.

Concerning the results of the fourth experiment, which examined the different GAN objectives, it is noticeable that WGAN-GP struggles to catch CT specific features compared to LSGAN. First of all, although WGGAN-GP shows an improvement in performance for pix2pix, the CycleGAN and CUT models both have difficulties. One could explain this by the higher computational complexity due to the additional loss functions in the architectures that perform the training in unpaired fashion. Due to the poor performance of WGAN models, the review of related work was performed. It was found that training Wasserstein GANs with momentum-based optimisers such as Adam, which was used in the default NN configuration for all experiments, becomes unstable. This is justified because the criterion loss is highly unsteady, which meant that momentum-based optimisers seemed to perform more poorly (Hunter, 2018). Another run of the experiment is required to change the optimiser and tune the gradient penalty weights.

In the fifth experiment, the role of the perceptual loss function in the generator was evaluated. A negligible reduction in geometric metrics was observed when L1 was replaced by the VGG-19-based loss function. This could be due to the fact that the VGG19 used in the experiment was trained on the ImageNet dataset, which was designed for use in visual object recognition research and not for medical image analysis. However, the presence of blur, which is sometimes forgivable in computer vision tasks, could be the reason for significant dosimetric differences. Solutions to the current limitations include additional optimisation of the loss weight λ, training the network on medical images, or introducing a different type of perceptual loss as suggested in the research by (Hiasa et al., 2018; Lei et al., 2019; Kang et al., 2021)

Finally, the performance of the GAN architectures specifically selected for research could be strongly influenced by the loss function and the combination of the different layers that compose

the generator and discriminator (Spadea et al., 2021). Regarding the DL models used in research for generating sCTs, the problem that has not yet been solved is the quality of bone restoration and, in particular, the generation of ribs, which may be among the organs at risk if the tumour is located close to them. The proposed attention-gate model for medical imaging as one of the modifications of the U-net, which automatically learns to focus on target structures of different shapes and sizes, especially by highlighting some silent features, seems to be feasible for solving this problem **?**. This model has been successfully applied to the large CT abdominal datasets for multi-class image segmentation, which deserve to be evaluated for the possibility of establishing reliable quality of DL-based sCT generation and its potential clinical appliance.

Chapter 7

Conclusion and Outlook

In the thesis, we have addressed the non-trivial question of synthetic CT generation for MRgRT with the particular focus on the sparsely studied area of the abdomen. The CUT architecture, trained in an unpaired fashion, was applied for the first time for the sCT generation task among with other state-of-the-art architectures, such as CycleGAN and Pix2pix. The models were trained on the clinical dataset, consisting of 76 patients, that was prepared for the purposes of this study. Optimisation of the various parameters, such as different preprocessing techniques (including the novel biologically motivated N-peaks normalisation), networks configuration and training objectives, helped to achieve generation of the high quality synthetic CT images, which were evaluated using geometric and dosimetric accuracy metrics. The proposed DL-based sCT generation methods may be considered clinically applicable for treatment planning in the abdominal area, with the mean DVH indicator discrepancies with the original dCT-based plan of less than 1% for all models and less than 0.5% for all PTV DVH indicators for the best performing model. The best performing model proposed in this study (CycleGAN) outperforms the state-of-the-arts for all DVH indicators, having PTV D_{mean} difference with original plan within 0.32%, as well as for most geometrical metrics. Generating the bone tissue with high accuracy remains a problem. Future work includes extending the applied network configuration by searching for the synergistic effect over the best found parameters, incorporating additional study data for unpaired training, and applying an attention-based model to generate sCT for the abdomen and other anatomical locations.

Appendix A

Attachements

Listing A.1: Pix2Pix Model default configuration

```
---------------- Options ---------------
           batch_size: 1
                beta1: 0.5
                beta2: 0.999
       checkpoints_dir: ./checkpoints
       continue_train: False
           crop_size: 256
         dataset_mode: aligned
            direction: BtoA  [default: AtoB]
         display_env: main
         display_freq: 400
           display_id: None
        display_ncols: 4
         display_port: 8097
       display_server: http://localhost
      display_winsize: 256
           easy_label: experiment_name
                epoch: latest
          epoch_count: 1
       evaluation_freq: 5000
            gan_mode: vanilla
             gpu_ids: 0
           init_gain: 0.02
           init_type: xavier
            input_nc: 1  [default: 3]
             isTrain: True  [default: None]
           lambda_L1: 100.0
                lr: 0.0002
        lr_decay_iters: 50
           lr_policy: linear
      max_dataset_size: inf
               model: pix2pix  [default: cut]
            n_epochs: 100  [default: 200]
        n_epochs_decay: 0  [default: 200]
```

```
           n_layers_D: 3
                name: nifti_pix2pix_lsgan_2d_baseline [default: experiment_name]
                 ndf: 64
                netD: basic
                netG: unet_256
                 ngf: 64
         no_antialias: False
       no_antialias_up: False
           no_dropout: True
              no_flip: False
              no_html: False
                 norm: batch
               normD: instance
               normG: instance
          num_threads: 4
           output_nc: 1  [default: 3]
              phase: train
           pool_size: 0
          preprocess: none  [default: resize_and_crop]
       pretrained_name: None
           print_freq: 100
      random_scale_max: 3.0
         save_by_iter: False
       save_epoch_freq: 5
      save_latest_freq: 5000
        serial_batches: False
stylegan2_G_num_downsampling: 1
              suffix:
      update_html_freq: 1000
             verbose: False
----------------- End -------------------
dataset [AlignedDataset]
model [Pix2PixModel]
optimiser [Adam]
```

Listing A.2: CycleGAN Model default configuration

```
----------------- Options ---------------
           batch_size: 1
                beta1: 0.5
                beta2: 0.999
       checkpoints_dir: ./checkpoints
       continue_train: False
            crop_size: 256
         dataset_mode: unaligned
            direction: AtoB
         display_env: main
         display_freq: 400
           display_id: None
```

```
  display_ncols: 4
   display_port: 8097
 display_server: http://localhost
display_winsize: 256
    easy_label: experiment_name
         epoch: latest
   epoch_count: 1
evaluation_freq: 5000
      gan_mode: lsgan
       gpu_ids: 0
     init_gain: 0.02
     init_type: xavier
      input_nc: 1  [default: 3]
       isTrain: True  [default: None]
      lambda_A: 10.0
      lambda_B: 10.0
lambda_identity: 0.5
           lr: 0.0002
 lr_decay_iters: 50
     lr_policy: linear
max_dataset_size: inf
        model: cycle_gan  [default: cut]
     n_epochs: 100  [default: 200]
 n_epochs_decay: 0  [default: 200]
    n_layers_D: 3
          ndf: 64
         netD: basic
         netG: resnet_9blocks
          ngf: 64
  no_antialias: False
no_antialias_up: False
    no_dropout: True
       no_flip: False
       no_html: False
         normD: instance
         normG: instance
   num_threads: 4
     output_nc: 1  [default: 3]
         phase: train
     pool_size: 50
    preprocess: none  [default: resize_and_crop]
pretrained_name: None
    print_freq: 100
random_scale_max: 3.0
   save_by_iter: False
save_epoch_freq: 5
save_latest_freq: 5000
 serial_batches: False
```

```
stylegan2_G_num_downsampling: 1
              suffix:
       update_html_freq: 1000
             verbose: False
----------------- End -------------------
dataset [UnalignedDataset]
model [CycleGANModel]
optimiser [Adam]
```

Listing A.3: CUT Model default configuration

```
 ----------------- Options ---------------
            CUT_mode: CUT
           batch_size: 1
               beta1: 0.5
               beta2: 0.999
        checkpoints_dir: ./checkpoints
         continue_train: False
             crop_size: 256
           dataset_mode: unaligned
              direction: AtoB
            display_env: main
           display_freq: 400
             display_id: None
          display_ncols: 4
           display_port: 8097
         display_server: http://localhost
        display_winsize: 256
            easy_label: experiment_name
                epoch: latest
           epoch_count: 1
        evaluation_freq: 5000
      flip_equivariance: False
              gan_mode: lsgan
               gpu_ids: 0
             init_gain: 0.02
             init_type: xavier
              input_nc: 1  [default: 3]
               isTrain: True  [default: None]
            lambda_GAN: 1.0
            lambda_NCE: 1.0
                  lr: 0.0002
          lr_decay_iters: 50
              lr_policy: linear
       max_dataset_size: inf
                model: cut
             n_epochs: 100  [default: 200]
         n_epochs_decay: 0  [default: 200]
            n_layers_D: 3
```

```
               nce_T: 0.07
             nce_idt: True
nce_includes_all_negatives_from_minibatch: False
           nce_layers: 0,4,8,12,16
                ndf: 64
               netD: basic
               netF: mlp_sample
             netF_nc: 256
               netG: resnet_9blocks
                ngf: 64
          no_antialias: False
        no_antialias_up: False
            no_dropout: True
              no_flip: False
              no_html: False
                normD: instance
                normG: instance
           num_patches: 256
           num_threads: 4
            output_nc: 1  [default: 3]
               phase: train
            pool_size: 0
            preprocess: none  [default: resize_and_crop]
        pretrained_name: None
            print_freq: 100
       random_scale_max: 3.0
          save_by_iter: False
        save_epoch_freq: 5
       save_latest_freq: 5000
         serial_batches: False
stylegan2_G_num_downsampling: 1
              suffix:
       update_html_freq: 1000
             verbose: False
----------------- End -------------------
dataset [UnalignedDataset]
model [CUTModel]
optimiser [Adam]
```

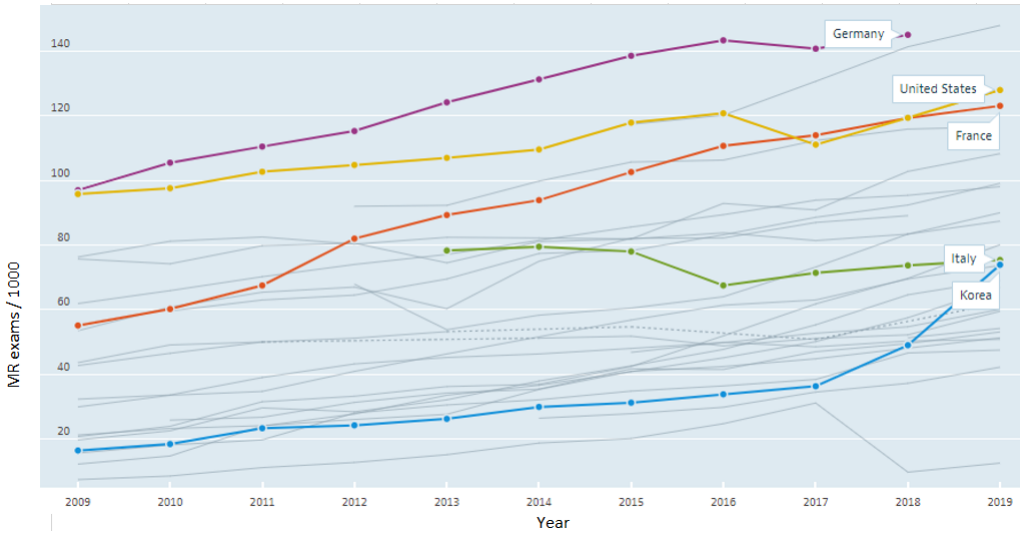


Figure A.1: MRI exams Total Per 1 000 inhabitants, 2009 – 2019 [https://data.oecd.org/]

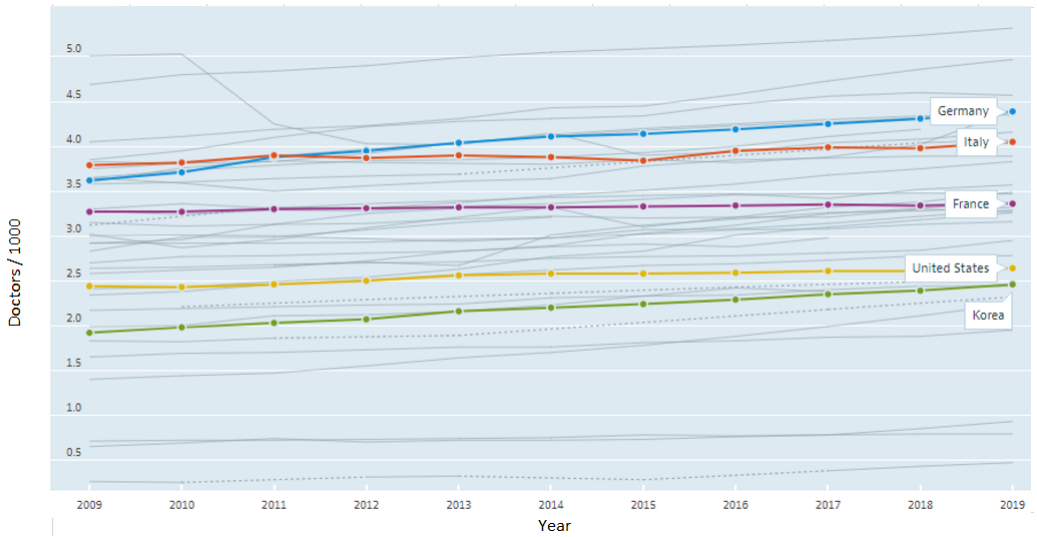


Figure A.2: Doctors Total Per 1 000 inhabitants, 2009 – 2019 [https://data.oecd.org/]

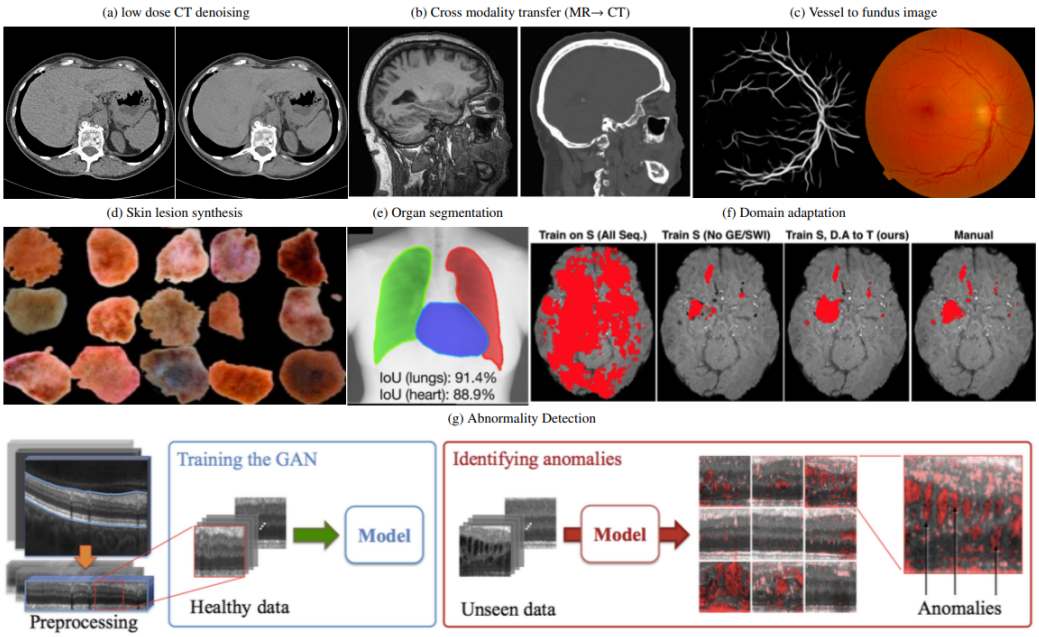


Figure A.3: Example applications using GANs. (a) Left side shows the noise contaminated low dose CT and right side shows the denoised CT that well preserved the low contrast regions in the liver Yi and Babyn (2018). (b) Left side shows the MR image and right side shows the synthesized corresponding CT. Bone structures were well delineated in the generated CT image Wolterink et al. (2017). (c) The generated retinal fundus image have the exact vessel structures as depicted in the left vessel map Costa et al. (2017). (d) Randomly generated skin lesion from random noise Yi et al. (2019). (e) An organ (lung and heart) segmentation example on adult chest X-ray. The shapes of lung and heart are regulated by the adversarial loss Dai et al. (2018). (f) The third column shows the domain adapted brain lesion segmentation result on SWI sequence without training with the corresponding manual annotation Kamnitsas et al. (2017). (g) Abnormality detection of optical coherence tomography images of the retina Schlegl et al. (2017). Source: Yi et al. Yi et al. (2019)

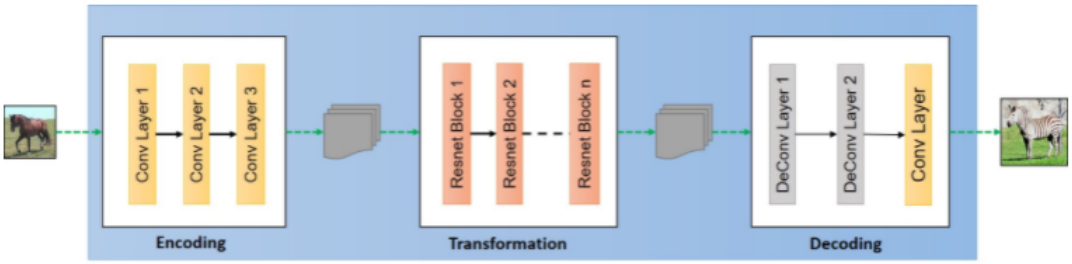


Figure A.4: Schematic representation of the CycleGAN ResNet-based generator. In contrast to the U-Net configuration, the ResNet has a "flatter" architecture, as the skip connections are retained in the transformation part. The first step, encoding, consists of extracting features from an image, which is done using a convolutional network. The goal of the transformation is to retain the features of the original input, such as the size and shape of the object, so ResNet is well suited for this type of transformation. Decoding is similar to the U-net and aims to reproduce the image in the same size.Source: CycleGAN blog Hardik Bansal (2017)

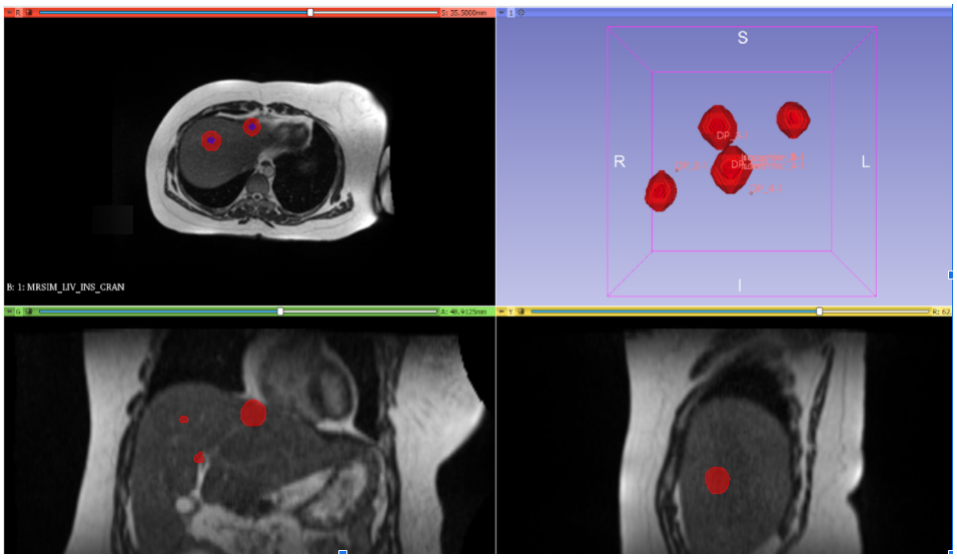


Figure A.5: Example of a patient with multiple tumours. PTV highlighted in red

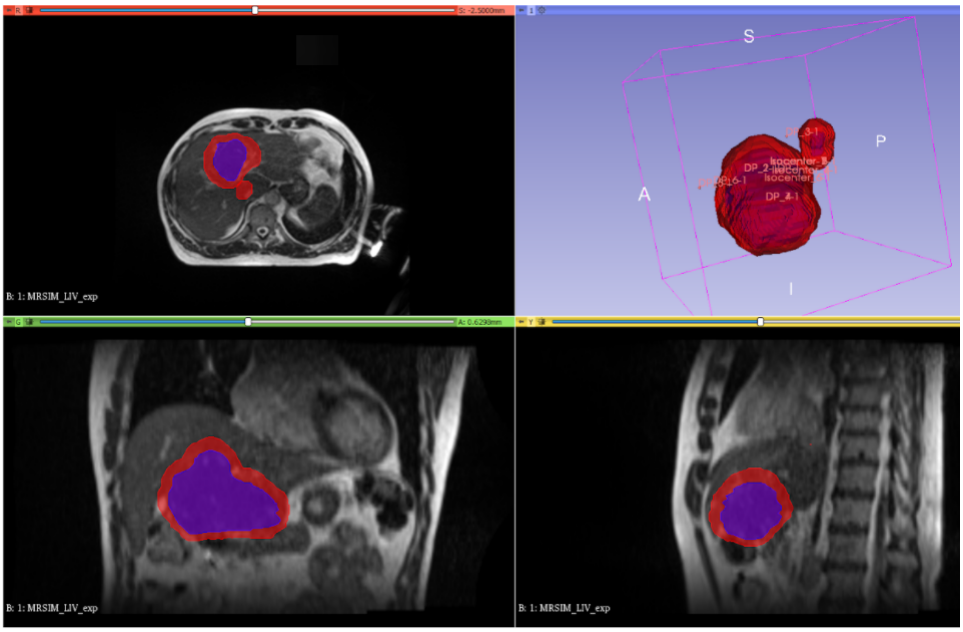
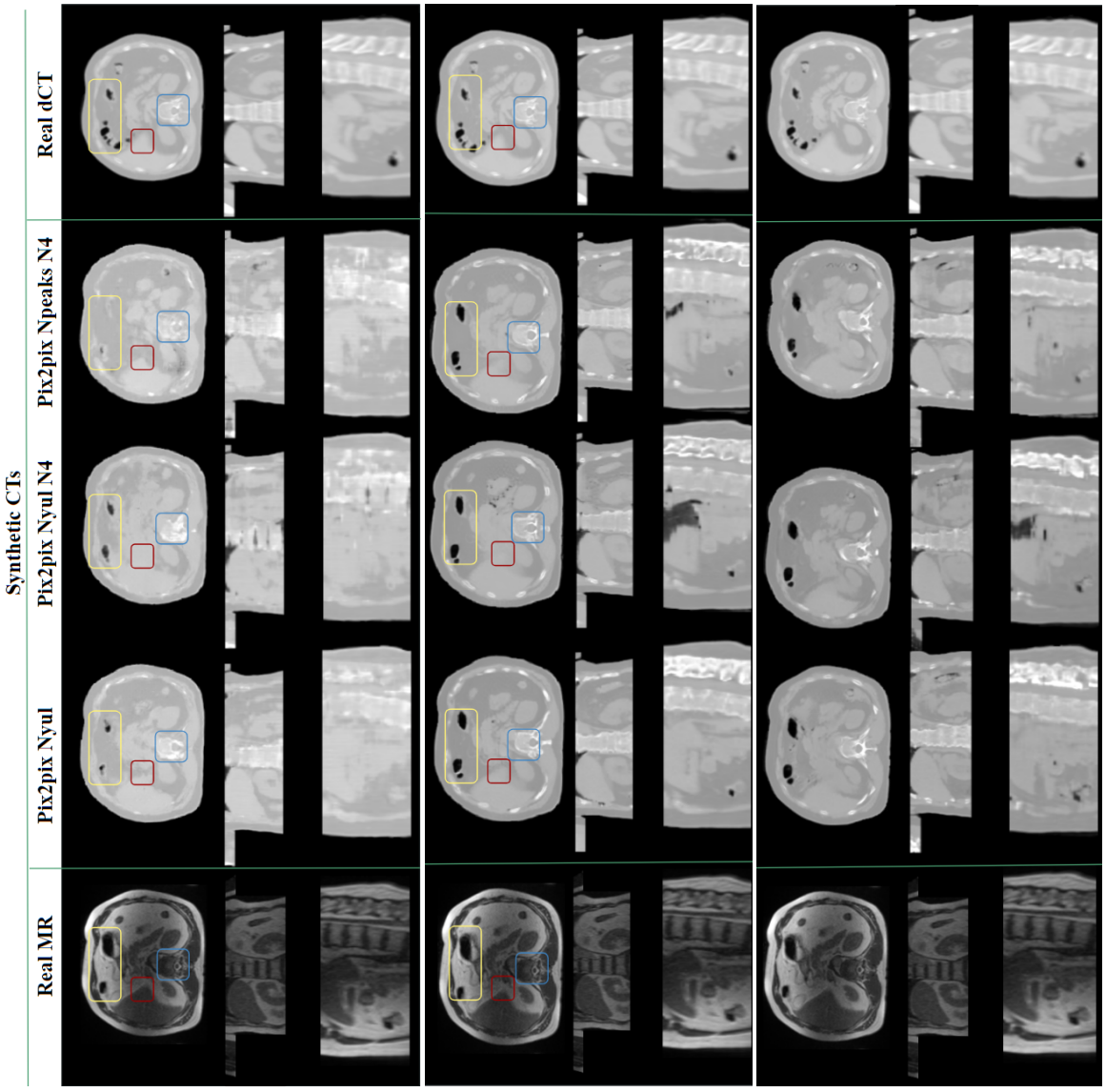


Figure A.6: Example of a patient with the largest tumour among the patients selected for the study. PTV highlighted in red



(a) pix2pix (b) CycleGAN (c) CUT

Figure A.7: EXPERIMENT 2. From left to right: original MR image; sCT, CycleGAN-generated synthetic CT, CUT-generated synthetic CT (all - fine-tuned); original de- formed computed tomography for a patient (high MAE case). From top to bottom: axial, coronal, sagital views. The rectangles highlight some areas of interest for reconstruction quality: yellow - air pockets, blue - ribs, red - liver edge

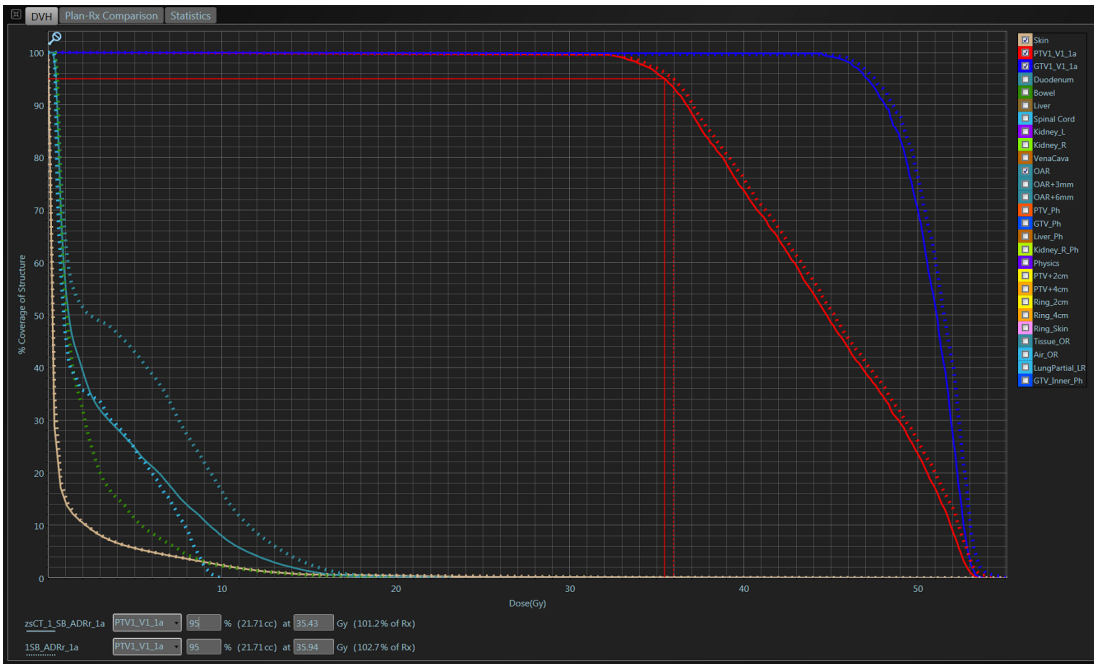


Figure A.8: Examples of the high DVH differences in sCT-based dosimetric evaluation. Solid lines shows the original dCT plan on DVH. Dotted line shows the sCT-based plan. Red lines shows the dose estimation for PTV

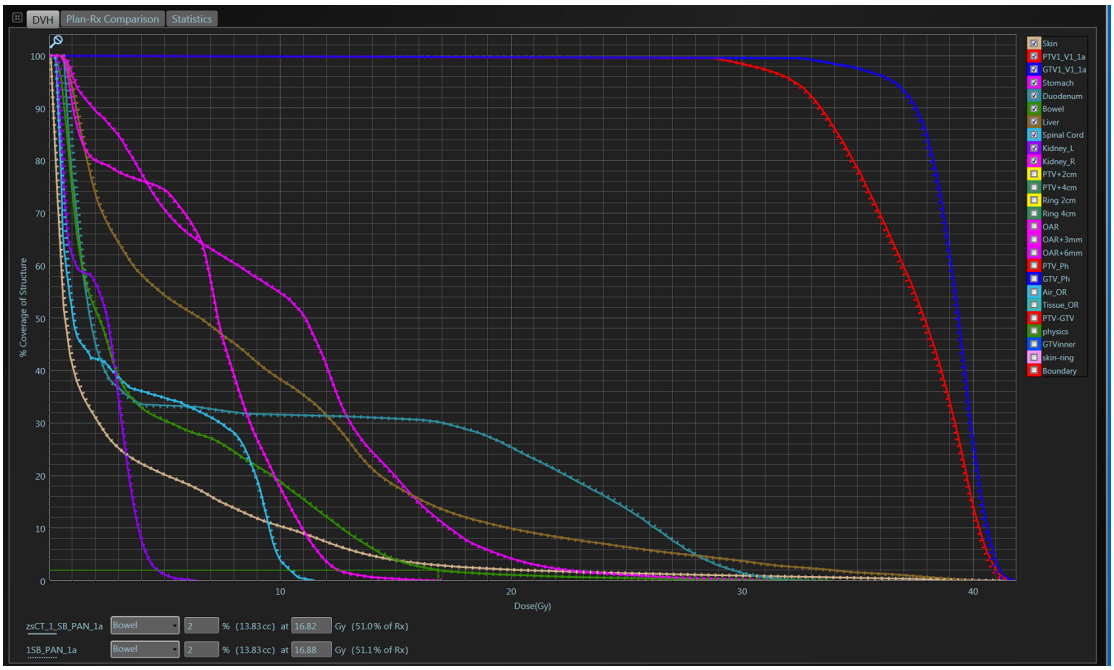


Figure A.9: Examples of the low DVH differences in sCT-based dosimetric evaluation. Solid lines shows the original dCT plan on DVH. Dotted line shows the sCT-based plan. Red lines shows the dose estimation for PTV

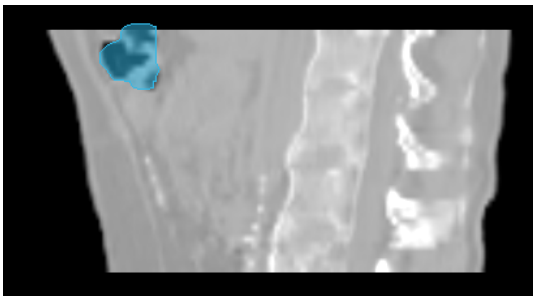


Figure A.10: The example of the sCT generated by CycleGAN, overlaid with the air pocket deliniated on original MR (AIR_OR structure) drawn in blue. The position of the air pocket is well captured by the DL-based method for sCT generation

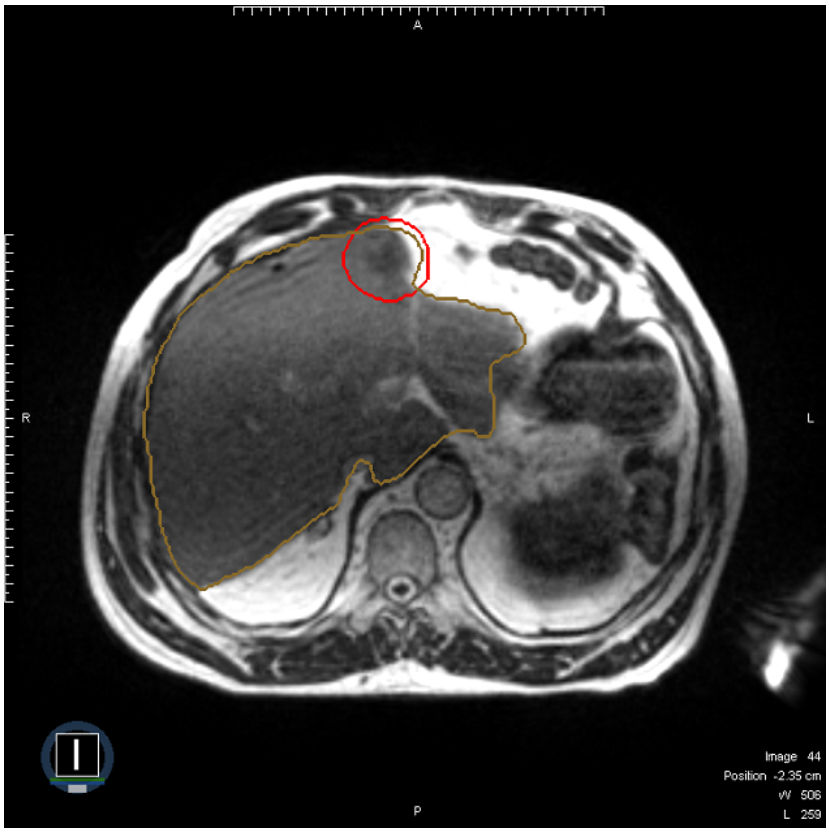


Figure A.11: Example of a low quality MR input (the presence of stripes along the entire liver, including the area close to the tumour, highlighted in red) affecting the formation of the generated sCT and leading to a difference in DVH parameters exceeding 1%

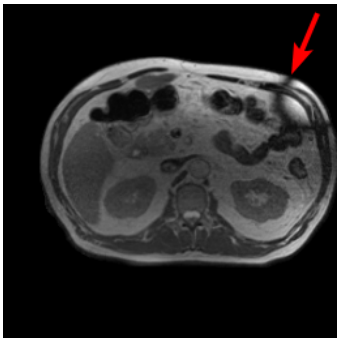


Figure A.12: Example of a low quality MR input (red arrow shows the artefact of the acquisition)

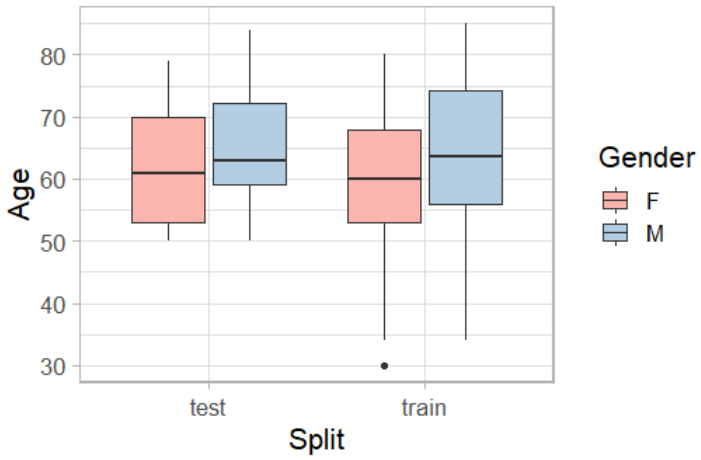


Figure A.13: Train and test set separation, represented by age and gender

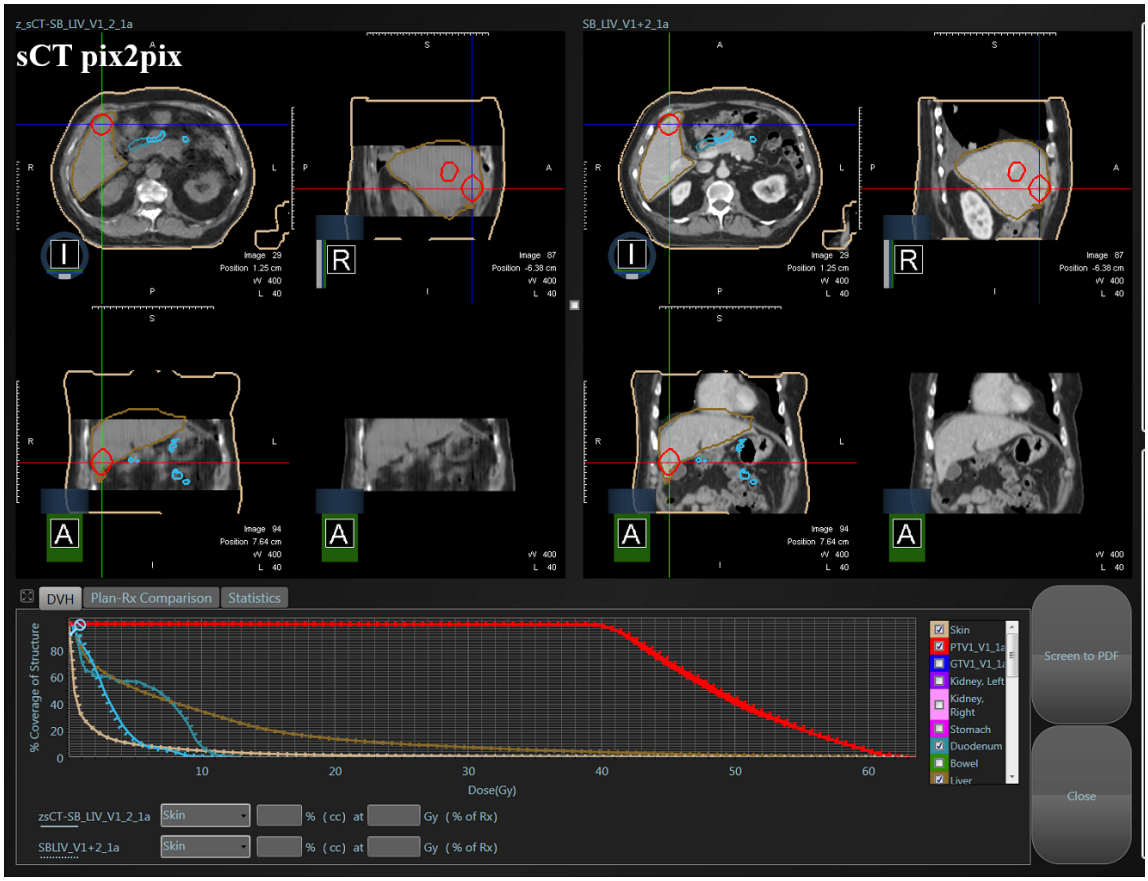


Figure A.14: Comparison of dose volume histogram for PTV and OARs between sCT generated by pix2pix (low MAE case) shown on left and original dCT shown on right. Dashed line indicates the result of dCT and the solid line shows the result of sCT methods

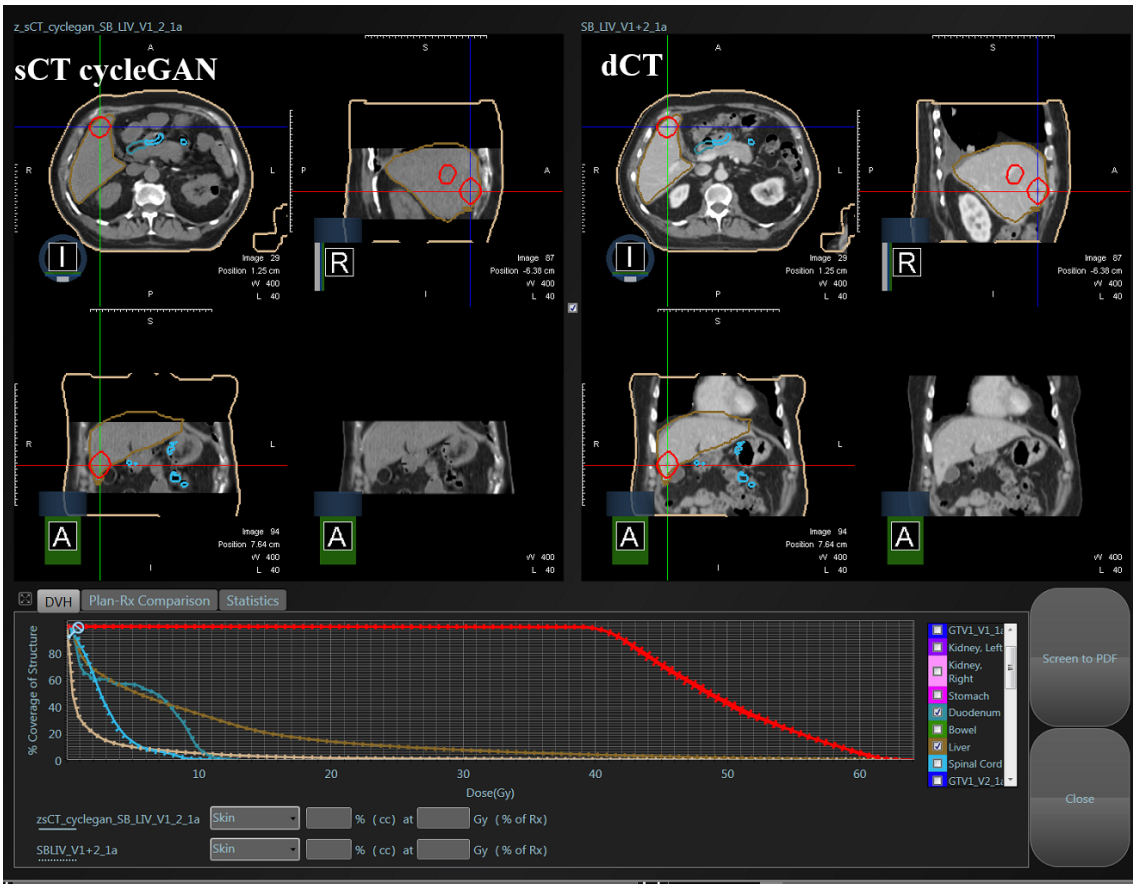


Figure A.15: Comparison of dose volume histogram for PTV and OARs between sCT generated by CycleGAN (low MAE case) shown on left and original dCT shown on right. Dashed line indicates the result of dCT and the solid line shows the result of sCT methods

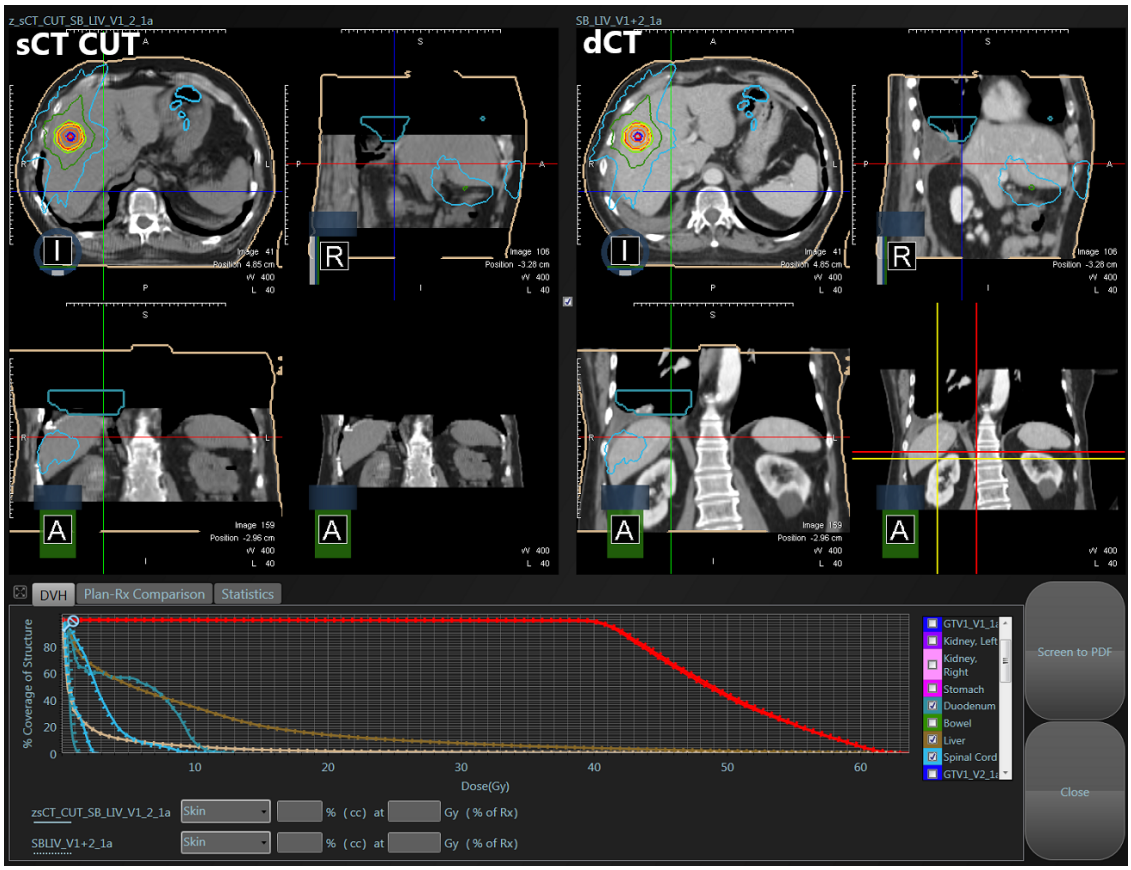


Figure A.16: Comparison of dose volume histogram for PTV and OARs between sCT generated by CUT (low MAE case) shown on left and original dCT shown on right. Dashed line indicates the result of dCT and the solid line shows the result of sCT methods

List of Figures

List of Tables

List of Listings

Bibliography

Ahad, R. (2015). *Development of Brain Computer Interface (BCI) System for Integration with Functional Electrical Stimulation (FES) Application*. PhD thesis, Universiti Tun Hussein Onn Malaysia.

Andres, E. A., Fidon, L., Vakalopoulou, M., Lerousseau, M., Carré, A., Sun, R., Klausner, G., Ammari, S., Benzazon, N., Reuzé, S., et al. (2020). Dosimetry-driven quality measure of brain pseudo computed tomography generated from deep learning for mri-only radiation therapy treatment planning. *International Journal of Radiation Oncology* Biology* Physics*, 108(3):813–823.

Antolak, J. A. and Rosen, I. I. (1999). Planning target volumes for radiotherapy: how much margin is needed? *International Journal of Radiation Oncology* Biology* Physics*, 44(5):1165–1170.

Arabi, H., Dowling, J. A., Burgos, N., Han, X., Greer, P. B., Koutsouvelis, N., and Zaidi, H. (2018). Comparative study of algorithms for synthetic ct generation from mri: consequences for mri-guided radiation planning in the pelvic region. *Medical physics*, 45(11):5218–5233.

Arjovsky, M., Chintala, S., and Bottou, L. (2017). Wasserstein generative adversarial networks. In *International conference on machine learning*, pages 214–223. PMLR.

Baskar, R., Lee, K. A., Yeo, R., and Yeoh, K.-W. (2012). Cancer and radiation therapy: current advances and future directions. *International journal of medical sciences*, 9(3):193.

Boulanger, M., Nunes, J.-C., Chourak, H., Largent, A., Tahri, S., Acosta, O., De Crevoisier, R., Lafond, C., and Barateau, A. (2021). Deep learning methods to generate synthetic ct from mri in radiotherapy: A literature review. *Physica Medica*, 89:265–281.

Broder, J. and Preston, R. (2011). Imaging the head and brain. *Diagnostic imaging for the emergency physician. WB Saunders*, pages 1–45.

Burgos, N. (2017). Iterative framework for the joint segmentation and ct synthesis of mr images burgos, ninon; guerreiro, filipa; mcclelland, jamie; presles, benoit; modat, marc; nill, simeon; dearnaley, david; desouza, nandita; oelfke, uwe; knopf, antje-christin. *Phys. Med. Biol*, 62:4237.

Chappell, M. (2019). *Principles of Medical Imaging for Engineers*. Springer.

Christensen, A. N., Larsen, C. T., Mandrup, C. M., Petersen, M. B., Larsen, R., Conradsen, K., and Dahl, V. A. (2017). Automatic segmentation of abdominal fat in mri-scans, using graph-cuts and image derived energies. In *Scandinavian Conference on Image Analysis*, pages 109–120. Springer.

Costa, P., Galdran, A., Meyer, M. I., Niemeijer, M., Abràmoff, M., Mendonça, A. M., and Campilho, A. (2017). End-to-end adversarial retinal image synthesis. *IEEE transactions on medical imaging*, 37(3):781–791.

Cusumano, D., Lenkowicz, J., Votta, C., Boldrini, L., Placidi, L., Catucci, F., Dinapoli, N., Antonelli, M. V., Romano, A., De Luca, V., et al. (2020). A deep learning approach to generate synthetic ct in low field mr-guided adaptive radiotherapy for abdominal and pelvic cases. *Radiotherapy and Oncology*, 153:205–212.

Dai, W., Dong, N., Wang, Z., Liang, X., Zhang, H., and Xing, E. P. (2018). Scan: Structure correcting adversarial network for organ segmentation in chest x-rays. In *Deep learning in medical image analysis and multimodal learning for clinical decision support*, pages 263–273. Springer.

Demol, B., Boydev, C., Korhonen, J., and Reynaert, N. (2016). Dosimetric characterization of mri-only treatment planning for brain tumors in atlas-based pseudo-ct images generated from standard t1-weighted mr images. *Medical physics*, 43(12):6557–6568.

Drzymala, R., Mohan, R., Brewster, L., Chu, J., Goitein, M., Harms, W., and Urie, M. (1991). Dose-volume histograms. *International Journal of Radiation Oncology* Biology* Physics*, 21(1):71–78.

Edmund, J. M. and Nyholm, T. (2017). A review of substitute ct generation for mri-only radiation therapy. *Radiation Oncology*, 12(1):1–15.

Estrada, S., Lu, R., Conjeti, S., Orozco-Ruiz, X., Panos-Willuhn, J., Breteler, M. M., and Reuter, M. (2020). Fatsegnet: A fully automated deep learning pipeline for adipose tissue segmentation on abdominal dixon mri. *Magnetic resonance in medicine*, 83(4):1471–1483.

Fard, A. S., Reutens, D. C., and Vegh, V. (2021). Cnns and gans in mri-based cross-modality medical image estimation. *arXiv preprint arXiv:2106.02198*.

Farjam, R., Nagar, H., Kathy Zhou, X., Ouellette, D., Chiara Formenti, S., and DeWyngaert, J. K. (2021). Deep learning-based synthetic ct generation for mr-only radiotherapy of prostate cancer patients with 0.35 t mri linear accelerator. *Journal of Applied Clinical Medical Physics*, 22(8):93–104.

Farjam, R., Tyagi, N., Deasy, J. O., and Hunt, M. A. (2019). Dosimetric evaluation of an atlas-based synthetic ct generation approach for mr-only radiotherapy of pelvis anatomy. *Journal of Applied Clinical Medical Physics*, 20(1):101–109.

Goodfellow, I. (2016). Nips 2016 tutorial: Generative adversarial networks. *arXiv preprint arXiv:1701.00160*.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial nets. *Advances in neural information processing systems*, 27.

Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., and Courville, A. C. (2017). Improved training of wasserstein gans. *Advances in neural information processing systems*, 30.

Han, J., Shoeiby, M., Petersson, L., and Armin, M. A. (2021). Dual contrastive learning for unsupervised image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 746–755.

Han, X. (2017). Mr-based synthetic ct generation using a deep convolutional neural network method. *Medical physics*, 44(4):1408–1419.

Hardik Bansal, A. R. (2017). Understanding and implementing cyclegan in tensorflow. Accessed at: https://hardikbansal.github.io/CycleGANBlog/.

He, K., Fan, H., Wu, Y., Xie, S., and Girshick, R. (2020). Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., and Hochreiter, S. (2017). Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30.

Hiasa, Y., Otake, Y., Takao, M., Matsuoka, T., Takashima, K., Carass, A., Prince, J. L., Sugano, N., and Sato, Y. (2018). Cross-modality image synthesis from unpaired data using cyclegan. In *International workshop on simulation and synthesis in medical imaging*, pages 31–41. Springer.

Hori, M., Hagiwara, A., Goto, M., Wada, A., and Aoki, S. (2021). Low-field magnetic resonance imaging: its history and renaissance. *Investigative Radiology*, 56(11):669.

Hou, K.-Y., Lu, H.-Y., and Yang, C.-C. (2021). Applying mri intensity normalization on non-bone tissues to facilitate pseudo-ct synthesis from mri. *Diagnostics*, 11(5):816.

Hunter, H. (2018). GAN Objective Functions: GANs and Their Variations. Accessed at: https://towardsdatascience.com/gan-objective-functions-gans-and-their-variations-ad77340bce3c.

Hussain, A. and Muhammad, W. (2017a). *Treatment Planning in Radiation Therapy*, pages 63–129.

Hussain, A. and Muhammad, W. (2017b). Treatment planning in radiation therapy. In *An Introduction to Medical Physics*, pages 63–129. Springer.

Iglesias, J. E. and Sabuncu, M. R. (2015). Multi-atlas segmentation of biomedical images: a survey. *Medical image analysis*, 24(1):205–219.

Isola, P., Zhu, J.-Y., Zhou, T., and Efros, A. A. (2017). Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134.

Johnstone, E., Wyatt, J. J., Henry, A. M., Short, S. C., Sebag-Montefiore, D., Murray, L., Kelly, C. G., McCallum, H. M., and Speight, R. (2018). Systematic review of synthetic computed tomography generation methodologies for use in magnetic resonance imaging–only radiation therapy. *International Journal of Radiation Oncology* Biology* Physics*, 100(1):199–217.

Kamalian, S., Lev, M. H., and Gupta, R. (2016). Computed tomography imaging and angiography–principles. *Handbook of clinical neurology*, 135:3–20.

Kamnitsas, K., Baumgartner, C., Ledig, C., Newcombe, V., Simpson, J., Kane, A., Menon, D., Nori, A., Criminisi, A., Rueckert, D., et al. (2017). Unsupervised domain adaptation in brain lesion segmentation with adversarial networks. In *International conference on information processing in medical imaging*, pages 597–609. Springer.

Kang, K. M., Choi, H. S., Jeong, B. K., Song, J. H., Ha, I.-B., Lee, Y. H., Kim, C. H., and Jeong, H. (2017). Mri-based radiotherapy planning method using rigid image registration technique combined with outer body correction scheme: a feasibility study. *Oncotarget*, 8(33):54497.

Kang, S. K., An, H. J., Jin, H., Kim, J.-i., Chie, E. K., Park, J. M., and Lee, J. S. (2021). Synthetic ct generation from weakly paired mr images using cycle-consistent gan for mr-guided radiotherapy. *Biomedical Engineering Letters*, 11(3):263–271.

Kim, K. (2018). Wgan. Accessed at: https://kionkim.github.io/2018/06/01/WGAN_1/.

Kim, S. I. and Suh, T. S. (2007). *World Congress of Medical Physics and Biomedical Engineering 2006: August 27-Septmber 1, 20006 COEX Seoul, Korea*, volume 14. Springer Science & Business Media.

Klages, P., Benslimane, I., Riyahi, S., Jiang, J., Hunt, M., Deasy, J. O., Veeraraghavan, H., and Tyagi, N. (2020). Patch-based generative adversarial neural network models for head and neck mr-only planning. *Medical physics*, 47(2):626–642.

Klüter, S. (2019). Technical design and concept of a 0.35 t mr-linac. *Clinical and Translational Radiation Oncology*, 18:98–101.

Konar, M. and Lang, J. (2011). Pros and cons of low-field magnetic resonance imaging in veterinary practice. *Veterinary radiology & ultrasound*, 52:S5–S14.

Largent, A., Barateau, A., Nunes, J.-C., Mylona, E., Castelli, J., Lafond, C., Greer, P. B., Dowling, J. A., Baxter, J., Saint-Jalmes, H., et al. (2019). Comparison of deep learning-based and patch-based methods for pseudo-ct generation in mri-based prostate dose planning. *International Journal of Radiation Oncology* Biology* Physics*, 105(5):1137–1150.

Lei, Y., Harms, J., Wang, T., Liu, Y., Shu, H.-K., Jani, A. B., Curran, W. J., Mao, H., Liu, T., and Yang, X. (2019). Mri-only based synthetic ct generation using dense cycle consistent generative adversarial networks. *Medical physics*, 46(8):3565–3581.

Li, W., Li, Y., Qin, W., Liang, X., Xu, J., Xiong, J., and Xie, Y. (2020). Magnetic resonance image (mri) synthesis from brain computed tomography (ct) images based on deep learning methods for magnetic resonance (mr)-guided radiotherapy. *Quantitative imaging in medicine and surgery*, 10(6):1223.

Library, N. (2011). Shielding radiation. alphas, betas, gammas and neutrons. Accessed at: https://www.nrc.gov/docs/ML1122/ML11229A721.pdf.

Mao, X., Li, Q., Xie, H., Lau, R. Y., Wang, Z., and Paul Smolley, S. (2017). Least squares generative adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2794–2802.

NEMA (2016). Dicom ps3.3 2016c - information object definitions. Accessed at: https://dicom.nema.org/medical/dicom/2016c/output/chtml/part03/sect_C.8.8.6.html.

NIBIB (2022). Magnetic Resonance Imaging (MRI). Accessed at: https://www.nibib.nih.gov/science-education/science-topics/magnetic-resonance-imaging-mri.

Nie, D., Trullo, R., Lian, J., Petitjean, C., Ruan, S., Wang, Q., and Shen, D. (2017). Medical image synthesis with context-aware generative adversarial networks. In *International conference on medical image computing and computer-assisted intervention*, pages 417–425. Springer.

Nyholm, T., Nyberg, M., Karlsson, M. G., and Karlsson, M. (2009). Systematisation of spatial uncertainties for comparison between a mr and a ct-based radiotherapy workflow for prostate treatments. *Radiation oncology*, 4(1):1–9.

Nyúl, L. G., Udupa, J. K., and Zhang, X. (2000). New variants of a method of mri scale standardization. *IEEE transactions on medical imaging*, 19(2):143–150.

Park, T., Efros, A. A., Zhang, R., and Zhu, J.-Y. (2020). Contrastive learning for unpaired image-to-image translation. In *European Conference on Computer Vision*, pages 319–345. Springer.

Park, T., Liu, M.-Y., Wang, T.-C., and Zhu, J.-Y. (2019). Semantic image synthesis with spatially-adaptive normalization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Pham, D. L., Xu, C., and Prince, J. L. (2000). Current methods in medical image segmentation. *Annual review of biomedical engineering*, 2(1):315–337.

Qin, S. and Jiang, T. (2018). Improved wasserstein conditional generative adversarial network speech enhancement. *EURASIP Journal on Wireless Communications and Networking*, 2018(1):1–10.

Radford, A., Metz, L., and Chintala, S. (2015). Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*.

Reinhold, J. C., Dewey, B. E., Carass, A., and Prince, J. L. (2019). Evaluating the impact of intensity normalization on MR image synthesis. In *Medical Imaging 2019: Image Processing*, volume 10949, page 109493H. International Society for Optics and Photonics.

Ronneberger, O., Fischer, P., and Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer.

Rosbergen, M. (2021). Ai & mri: improved mapping of oxygenation in brain tumours.

Schlegl, T., Seeböck, P., Waldstein, S. M., Schmidt-Erfurth, U., and Langs, G. (2017). Unsupervised anomaly detection with generative adversarial networks to guide marker discovery. In *International conference on information processing in medical imaging*, pages 146–157. Springer.

Seitzer, M. (2020). pytorch-fid: FID Score for PyTorch. Accessed at: https://github.com/mseitzer/pytorch-fid. Version 0.2.1.

Sharon and Eda Zhou, E. Z. (2020). Pix2pix overview. Accessed at: https://www.coursera.org/lecture/apply-generative-adversarial-networks-gans/pix2pix-overview-X9EOT.

Shinohara, R. T., Sweeney, E. M., Goldsmith, J., Shiee, N., Mateen, F. J., Calabresi, P. A., Jarso, S., Pham, D. L., Reich, D. S., Crainiceanu, C. M., et al. (2014). Statistical normalization techniques for magnetic resonance imaging. *NeuroImage: Clinical*, 6:9–19.

Shrivastava, A., Pfister, T., Tuzel, O., Susskind, J., Wang, W., and Webb, R. (2017). Learning from simulated and unsupervised images through adversarial training. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2107–2116.

Simonyan, K. and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Song, S., Zheng, Y., and He, Y. (2017). A review of methods for bias correction in medical images. *Biomedical Engineering Review*, 1(1).

Spadea, M. F., Maspero, M., Zaffino, P., and Seco, J. (2021). Deep learning based synthetic-ct generation in radiotherapy and pet: A review. *Medical physics*, 48(11):6537–6566.

Spindeldreier, C. K., Klüter, S., Hoegen, P., Buchele, C., Rippke, C., Tonndorf-Martini, E., Debus, J., and Hörner-Rieber, J. (2021). Mr-guided radiotherapy of moving targets. *Der Radiologe*, 61(1):39–48.

Sprawls, P. (2000). *Magnetic resonance imaging: principles, methods, and techniques*. Medical Physics Publishing Madison.

Taigman, Y., Polyak, A., and Wolf, L. (2016). Unsupervised cross-domain image generation. *arXiv preprint arXiv:1611.02200*.

Takano, N. and Alaghband, G. (2020). Generator from edges: Reconstruction of facial images. In *International Symposium on Visual Computing*, pages 430–443. Springer.

Thévenaz, P., Blu, T., and Unser, M. (2000). Image interpolation and resampling. *Handbook of medical imaging, processing and analysis*, 1(1):393–420.

Tustison, N. J., Avants, B. B., Cook, P. A., Zheng, Y., Egan, A., Yushkevich, P. A., and Gee, J. C. (2010). N4itk: improved n3 bias correction. *IEEE transactions on medical imaging*, 29(6):1310–1320.

Ulin, K., Urie, M. M., and Cherlow, J. M. (2010). Results of a multi-institutional benchmark test for cranial ct/mr image registration. *International Journal of Radiation Oncology* Biology* Physics*, 77(5):1584–1589.

Wallimann, P., Mayinger, M., Bogowicz, M., and Matthias Guckenberger, Nicolaus Andratschke, S. T.-L. J. E. v. T. H. G. (2022). N peaks normalisation. *Available by request: Philipp.Wallimann@usz.ch*.

Wang, T. and Lin, Y. (2018). Cyclegan with better cycles.

Whitcher, B., Schmid, V. J., and Thorton, A. (2011). Working with the dicom and nifti data standards in r. *Journal of Statistical Software*, 44:1–29.

WHO (2022). Cancer. key facts. Accessed at: https://www.who.int/news-room/fact-sheets/detail/cancer.

Wolterink, J. M., Dinkla, A. M., Savenije, M. H., Seevinck, P. R., van den Berg, C. A., and Išgum, I. (2017). Deep mr to ct synthesis using unpaired data. In *International workshop on simulation and synthesis in medical imaging*, pages 14–23. Springer.

Wyckoff, H. O., Allisy, A., and Lidén, K. (1976). The new special names of si units in the field of ionizing radiations. *The British Journal of Radiology*, 49(581):476–477.

Xu, K., Cao, J., Xia, K., Yang, H., Zhu, J., Wu, C., Jiang, Y., and Qian, P. (2019). Multichannel residual conditional gan-leveraged abdominal pseudo-ct generation via dixon mr images. *IEEE Access*, 7:163823–163830.

Yi, X. and Babyn, P. (2018). Sharpness-aware low-dose ct denoising using conditional generative adversarial network. *Journal of digital imaging*, 31(5):655–669.

Yi, X., Walia, E., and Babyn, P. (2019). Generative adversarial network in medical imaging: A review. *Medical image analysis*, 58:101552.

Zhu, J.-Y., Park, T., Isola, P., and Efros, A. A. (2017). Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232.